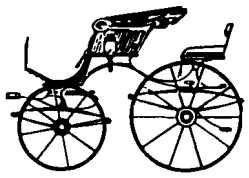


LABORATOIRE D'ECONOMIE DES TRANSPORTS
Ecole Nationale des Travaux Publics de l'Etat - Université Lumière Lyon II
Unité mixte de Recherche du CNRS n° 108.

PRODUCTION ET TRAITEMENT DES DONNEES
(Notes de Cours)

Philippe CALVIE
Economiste des Transports
BCEOM

Didier PLAT
Ingénieur TPE
Laboratoire d'Economie des Transports

Bernard SCHEOU
Attaché de recherche
Laboratoire d'Economie des Transports

Université d'été
Août 1995

INTRODUCTION

L'action pour transformer certains aspects d'une société, faire évoluer des pratiques professionnelles impose une connaissance du milieu sur lequel on souhaite agir. Connaître pour agir, connaître avant d'agir peuvent sembler des banalités. Mais force est de constater que le secteur transport, dans un certain nombre de pays en développement, reste largement terra incognita pour ceux-là mêmes qui en sont en charge au sein de l'administration ou chez les bailleurs de fonds internationaux. Or, cette carence est au mieux inopérante, au pire nuisible : si une connaissance partielle du réel est inévitable et doit être acceptée, sa connaissance partiale doit par contre être évitée. (ex. du transport informel dans beaucoup de pays, "la réalité se venge").

L'efficacité de l'action passe alors par une bonne connaissance des mécanismes de fonctionnement de la "réalité" que l'on essaie de maîtriser ou d'orienter. Le recours à la théorisation (représentation organisée de la réalité) doit permettre de dépasser le stade de l'action aveugle, mais implique des validations empiriques ou des confrontations régulières avec l'expérimentation.

Cette nécessité d'une articulation entre une théorisation et une appréhension empirique de la réalité achoppe, et tout particulièrement dans les pays en développement, sur un double handicap en matière de données. Elles sont en effet à la fois trop rares et trop peu utilisées. La rareté tient essentiellement au coût de leur production, mais aussi dans certains cas au peu d'intérêt porté au secteur transport. La sous-utilisation est liée également à la carence des moyens. En particulier, l'accès à l'outil informatique est souvent malaisé, soit par manque de matériel (ordinateurs ou logiciels), soit faute de personnel qualifié pour assurer les traitements et l'analyse.

Il est possible de distinguer deux manières d'appréhender la réalité. La première est ponctuelle et concerne des projets précis d'importance diverses comme le réaménagement d'une gare routière par exemple, ou la "grosse" enquête réalisée dans le cadre d'un plan national de transport. La seconde est permanente, ou du moins régulière et permet une connaissance du secteur des transport s'inscrivant dans la durée.

Les deux premières parties de ce module traiteront de ces deux formes de connaissance, la troisième ayant trait aux méthodes statistiques de traitement de l'information :

- la production "ponctuelle" d'informations,
- l'organisation d'un système permanent d'information statistique sur le secteur des transports,
- les méthode statistique de traitement des informations.

A - LES ENSEMBLES PONCTUELS DE DONNEES

Rappelons tout d'abord que ponctuel est entendu ici par opposition à permanent : les données considérées ici sont recueillies et traitées dans le but de satisfaire à l'analyse de problèmes spécifiques, leur production n'est pas répétitive ou la fréquence en est très faible. Il ne faut pas pour autant conclure à une homogénéité de ces ensembles de données ponctuels, les problèmes posés étant eux-mêmes très hétérogènes, de l'étude isolée sur tel ou tel aspect du fonctionnement du secteur à la conception d'un Plan national de transport.

La production de données transport, ne serait-ce qu'à cause de son coût, ne saurait être une fin en soi. Elle vise à apporter des réponses à des questions et doit donc permettre des traitements statistiques susceptibles de fonder ces réponses : ne pas relever l'âge des individus lors de l'enquête interdira ensuite toute évaluation de l'impact de ce facteur sur leur niveau de mobilité ! Enregistrer la distance parcourue en fonction de classes prédéfinies prohibera de même toute régression du prix de transport selon la distance.

De la production au traitement des données, il y a donc une chaîne aux maillons interdépendants et il importe de concevoir chacune des étapes en fonction de l'ensemble de la chaîne. En dépit de ces interrelations, on présentera successivement les deux premières phases de la chaîne, la conception des enquêtes et leur réalisation. La dernière phase, l'analyse des données, est commune aux ensembles ponctuels de données et aux systèmes d'information statistique et fait l'objet de la dernière partie.

1 - CONCEVOIR LES ENQUETES

Avant de présenter un essai de classification des principales enquêtes utilisées en transport et de rappeler un ensemble de définitions, nous insistons sur l'importance de la phase amont de détermination des besoins en données.

1.1 - DEFINIR LES BESOINS

Que l'on se situe dans une démarche opérationnelle (réponse aux demandes des politiques) ou dans une logique de recherche, on se trouve initialement confronté au même handicap d'imprécision des besoins. En effet, les questions posées au planificateur, à l'homme d'étude, au chercheur vont être le plus souvent formulées en termes très généraux : comment diminuer les prix de transport, que se passe-t-il si l'on augmente les taxes à l'importation, ... Une même exigence, une même nécessité s'imposent alors au spécialiste : arriver à sérier les questions, les préciser et les organiser au sein d'une vision globale.

Cette formalisation du problème posé implique la définition d'un certain nombre de concepts et la mise en évidence de leurs interrelations, c'est-à-dire la génération d'un cadre conceptuel, support et élément d'une théorisation. On va ainsi isoler de petits morceaux de réel que l'on va nommer, dont on va définir les propriétés et que l'on va agencer afin d'en décrire les interactions, les articulations. Mais cette opération de découpage va laisser de côté des chutes, des morceaux de réalité que l'on ignorera. Remarquons immédiatement deux conséquences de l'opération. D'une part, le résultat obtenu est une représentation de la réalité, pas la réalité elle-même : la théorie n'est pas la réalité que l'on observe, mais juste sa représentation. D'autre part, à travers le processus de sélection, on obtiendra une connaissance organisée mais partielle de la réalité : la théorie, non seulement n'est pas la réalité, mais n'en est qu'une représentation, une parmi d'autres.

Implicitement ou explicitement, consciemment ou inconsciemment, tout effort de théorisation et plus généralement toute opération sur le réel, va se trouver confronté à ce double impératif du choix et de l'organisation. Schématiquement, les choix vont déboucher sur la création de concepts, l'organisation sur la définition de leurs articulations. Certaines de ces articulations, de ces relations entre concepts (notamment des relations de cause à effet) pourront être élaborées sous la forme d'hypothèses que tant les développements ultérieurs de la théorie que les tests expérimentaux seront amenés ensuite à valider.

Ainsi, cette théorisation devra tout à la fois contribuer à préciser les besoins de l'étude en aidant à formuler et à affiner les questions envisagées et simultanément fournir le schéma adéquat duquel il faudra faire émerger les réponses. En permettant donc simultanément de poser les bonnes questions et de trouver les bonnes réponses, cette première phase dans le procesus de conception des enquêtes s'avère indispensable.

Deux remarques encore avant de présenter un essai de classification des enquêtes transport. D'une part, le fait que la théorisation ne débouche que sur une représentation parmi d'autres de la réalité n'implique nullement que ces diverses représentations sont équivalentes. En fonction du problème de départ, certaines, parce qu'elles négligent tel ou tel aspect ou postulent telle ou telle relation erronée, vont se cogner aux murs de la réalité. Leurs prédictions seront erronées, leurs hypothèses ne seront pas vérifiées. Mieux vaut alors essayer de changer la théorie plutôt que de changer la réalité ... D'autre part, cette attitude de réduction du champ embrassé afin de gagner en intelligibilité, nous le retrouverons à plusieurs reprises dans la suite de l'exposé. L'organisation et la compréhension de l'information passent bien souvent par sa réduction.

> *Le LET a réalisé en collaboration avec divers centres de recherche français et africains en économie des transports une recherche sur des politiques permettant de réduire les coûts du camionnage en Afrique subsaharienne (ASS). Lancée à l'initiative de la Banque mondiale mais sur financement de la Coopération française, elle visait à confirmer la réputation de cherté du camionnage en ASS et à en isoler les causes.*
> *Il a donc fallu préciser ce que l'on entendait par prix du camionnage (définition d'un concept), par coûts d'exploitation des entreprises (s'agit-il d'un coût réel - comptable- ou d'un coût apparent - perçu par le transporteur- ?), quelle pouvait être la relation entre prix et coûts (liaison entre deux concepts), ... Un cadre conceptuel a ainsi été élaboré.*
> *L'une des hypothèses de base était que le secteur des transports routiers de marchandises est fortement hiérarchisé, qu'aussi bien les prix que les coûts et que les relations entre prix et coûts varient d'un niveau hiérarchique à l'autre. Par hiérarchie (autre concept de base), on entend une relation forte entre les divers paramètres susceptibles de caractériser une expédition : quantité, distance, accompagnement de la marchandise, origine et destination, nature du contact entre chargeur et transporteur, ...*
> *Afin de valider cette hypothèse, de mesurer les prix et les coûts, de repérer des surcoûts, ... nous avons réalisé des relevés de prix et des entretiens sur les pratiques de transport et les postes de coût auprès des principaux acteurs du secteur dans trois pays : Cameroun, Côte d'ivoire, Mali.*
> *Les exemples (en italique) de cette première partie porteront sur ce travail.*

1.2 - LES ENQUETES TRANSPORT

Les techniques d'enquête, qu'elles concernent ou non le secteur transport, se ventilent en techniques qualitatives et techniques quantitatives. Idéalement, les premières visent à éclairer en profondeur un phénomène, à travers un ensemble d'entretiens réalisés auprès des principaux acteurs, alors que les secondes visent moins à décortiquer le problème qu'à en prendre la mesure. Mais cette présentation est trop caricaturale : les techniques quantitatives servent aussi à valider des schémas explicatifs tandis que les approches qualitatives peuvent fournir des éléments d'estimation. Dans la pratique, les oppositions sont d'ailleurs beaucoup moins marquées, il y a plutôt complémentarité entre les deux approches et parfois interpénétration.

Il existe différentes façons de classer les techniques quantitatives d'enquête utilisées dans le champ transport. On peut ainsi se fonder sur une grille par mode, par type spatial de trafic (urbain, non urbain), par nature des informations recensées (valeurs, comportements, perception des modes, ...), ... Sans chercher à être exhaustif dans le recensement des techniques, on s'appuiera ici sur une organisation par type spatial, en considérant toutefois à part les comptages.

LES COMPTAGES

Les comptages peuvent être utilisés aussi bien en milieu urbain qu'en rase campagne. Ils permettent de connaître les flux de trafic, soit pour un mode donné, soit pour l'ensemble des modes de transport. Ils sont ainsi utilisés pour des réaménagements de voirie (dimensionnement de chaussée par exemple), pour suivre l'évolution du trafic et de sa composition sur une portion du réseau d'infrastructures ou pour mesurer des vitesses d'écoulement des flux. Ils peuvent être utiles aussi dans le cadre de l'élaboration de matrices Origine-Destination (matrices OD) ou pour dimensionner des stratégies de régulation du trafic.

Ils peuvent être manuels ou automatiques, et dans ce dernier cas permanents ou temporaires. Ils peuvent être réalisés sur un tronçon donné d'infrastructure ou concerner l'ensemble des mouvements effectués par les véhicules à un carrefour, on parle alors de comptage directionnel.

LE MILIEU URBAIN

Toujours sans chercher à être parfaitement exhaustifs, on définira quatre grandes catégories d'enquêtes en milieu urbain.

Les enquêtes OD diffèrent selon le mode concerné. Pour les modes collectifs, la technique la plus usuelle consiste en l'utilisation de cartes à coupons détachables remises à chaque usager lors de sa montée dans le bus et récupérées lors de sa descente. Un code figurant sur la carte identifie l'entrée dans le réseau tandis que la sortie est repérée facilement dès lors que l'on prend soin de ne pas mélanger les paquets constitués à chacun des arrêts ; les coupons détachés correspondent aux réponses à des questions supplémentaires (par exemple le sexe, l'âge, ...). Les OD reconstituées ainsi sont toutefois de "fausses" OD (handicap des correspondances, assimilation des extrémités du déplacement aux arrêts du réseau TC). Pour les modes individuels (voiture, mais aussi deux roues), deux techniques s'affrontent, les enquêtes par identification du véhicule (relevés de plaques minéralogiques -sur papier, sur magnétophone ou même directement sur micro-ordinateur-, collage de papillons de couleur) et les enquêtes par interview, nécessitant l'arrêt du véhicule en bordure de voie. On conçoit aisément que les secondes puissent fournir des résultats plus élaborés (et notamment de véritables OD), même si les premières peuvent prétendre à la simplicité et à l'exhaustivité.

Des enquêtes spécifiques aux transports collectifs sont conçues pour mieux connaître la clientèle utilisant ces services. On distingue ainsi les comptages de montée et descente aux arrêts qui fournissent la fréquentation des stations et la charge des tronçons, les enquêtes embarquées qui contribuent à la reconstitution de matrices OD mais informent également les exploitants sur les besoins, représentations, récriminations ... des usagers.

Les enquêtes ménages concernent généralement un grand nombre de ménages interrogés à leur domicile sur leur mobilité (fréquence et motifs de déplacement, usage des modes, ...), leur connaissance et leur appréciation du réseau de transport collectif, ... Il s'agit là de procédures généralement très lourdes et très onéreuses, les informations recueillies étant toutefois plus riches.

Les enquêtes aux générateurs de trafic sont réalisées dans des centres d'activité importants (centres commerciaux, hôpitaux, grands lycées ou universités ...), le plus généralement à base de questionnaires administrés. Elles renseignent sur les pratiques de déplacement liées à ces générateurs et dans le cas de centres à rayonnement régional peuvent fournir des indications sur la mobilité non urbaine.

LE MILIEU INTERURBAIN

En plus bien sûr des divers types de comptage, on retrouve en milieu interurbain la plupart des techniques utilisées pour les transports urbains. C'est ainsi que des enquêtes auprès des ménages peuvent servir à reconstituer les comportements de déplacement non urbains ou que des enquêtes dans les cars ou dans les gares routières permettent de mieux connaître les usagers des transports collectifs.

Les pesées d'essieu sont par contre plus spécifiques de la rase campagne. Elles sont réalisées soit en arrêtant les véhicules et en pesant alors essieu par essieu (la balance peut alors être portable), soit par une balance dynamique insérée dans la chaussée (le poste de pesage est alors bien évidemment fixe). Le principal problème concerne bien évidemment la résistance et/ou l'entretien du matériel ... Ces pesées sont par contre le meilleur moyen d'obtenir des informations fiables sur la charge des véhicules.

1.3 - QUELQUES DEFINITIONS

Un questionnaire d'enquête est auto-administré dès lors que c'est l'enquêté qui le renseigne. Dans le cas où un enquêteur assure la passation du questionnaire, on parle de questionnaire (ou d'enquête) administrée. La première technique est bien évidemment moins onéreuse que la seconde. Mais les questionnaires complexes (par exemple ceux d'enquêtes ménages) nécessitent généralement la présence d'un enquêteur professionnel susceptible d'expliquer aux enquêtés, ou à certains d'entre eux, d'éventuels points obscurs. Notons toutefois que ces différences de traitement entre enquêtés risquent de conduire à l'apparition de biais.

La population est l'ensemble étudié. Elle se compose d'individus, éléments de base de cet ensemble. Il peut exister dans une même enquête plusieurs populations "emboîtées" les unes dans les autres. Des niveaux d'observation différents peuvent ainsi coexister dans une même enquête. On peut ainsi être amené à relever simultanément un ensemble de caractéristiques de ménages (première population) ainsi que des caractéristiques de chacun des individus faisant partie de ces ménages (seconde population). On parlera dans ce cas d'enquêtes hiérarchisées.

L'enquête camionnage est une enquête hiérarchisée où figurent plusieurs populations. En effet des informations (structure du parc par exemple) ont été recueillies auprès d'entreprises (première population) puis des données sur des expéditions récentes (de une à trois) ont été saisies dans chaque entreprise (les expéditions constituent la seconde population, située à un niveau hiérarchique inférieur).

Les analyses peuvent être menées à des niveaux d'observation, mais aussi s'appuyer sur les résultats de procédures d'agrégation et/ou désagrégation. C'est ainsi que des comparaisons internationales sur des budgets moyens, des volumes de mobilité, ... désignent des niveaux d'analyse (le pays en l'occurence) qui ne sont pas simultanément des niveaux d'observation. La production des données ne doit donc pas être cantonnée au seul travail de terrain. Nous reviendrons ultérieurement sur ces productions "dérivées" de données.

Le niveau hiérarchique ou le pays sont des niveaux d'analyse et non pas d'observation. Un certain nombre de données seront donc élaborées pour chacun de ces niveaux (par exemple, prix moyen unitaire par niveau hiérarchique)

Les individus sont observés à l'aide de caractères présentant deux ou plusieurs modalités. Les questions correspondantes sont dites fermées lorsque les modalités entre lesquelles l'enquêté choisit ont été définies préalablement, ouvertes lorsque la réponse est libre. Ces caractères renvoient aux concepts que l'on a retenus dans la phase initiale. Ils en sont une traduction, plus ou moins grossière, et donc une trahison.

Le concept de hiérarchie, tel qu'il a été défini, se présente comme un phénomène multidimensionnel complexe. Plusieurs stratégies permettent de l'approcher empiriquement : on peut situer une expédition dans la hiérarchie en se fondant uniquement sur la distance parcourue ou sur la quantité transportée ou bien en combinant ces deux paramètres ou bien en agrégeant trois, quatre, ... paramètres la caractérisant. Mais dans tous les cas, plusieurs dimensions du phénomène multidimensionnel "hiérarchie" seront occultées. L'indicateur empirique trahira ainsi le concept.

Les caractères, qui correspondent aux items du questionnaire, peuvent être de nature différente. On distingue des caractères qualitatifs (modalités non mesurables, l'ensemble des modalités constitue une nomenclature) et des caractères quantitatifs (soit discrets, c'est-à-dire à valeurs isolées ; soit continus, à valeurs en nombre infini). Il est toujours possible de générer un caractère qualitatif à partir d'une variable statistique en regroupant ses modalités en un nombre fini de classes. Enfin, dans le cas d'enquêtes hiérarchisées, certains caractères n'auront d'autre fonction que d'assurer une liaison, une filiation entre les différents niveaux (par exemple, un "numéro du ménage", figurant dans le questionnaire ménage et dans le questionnaire individu et prenant une valeur unique pour les différents membres d'un même ménage).

Le type de véhicule est un caractère qualitatif à trois modalités (utilitaire léger ou pick-up, camion porteur, ensemble articulé). Le tonnage exact, à supposer qu'il puisse être mesuré, est une variable quantitative continue. Arrondi à la tonne la plus proche, il devient une variable statistique discrète. Regroupé en quelques classes (par exemple moins de 3 tonnes, de 3 à 6, de 6 à 11, de 11 à 14.5, de 14.5 à 27 et plus de 27 tonnes), il constitue un caractère qualitatif Le numéro d'entreprise, numéro correspondant à l'ordre dans lequel les entreprises ont été enquêtées, est une variable de liaison, présente au niveau d'observation "entreprise" et au niveau d'observation "expédition".

L'enquête est exhaustive lorsqu'elle touche l'ensemble de la population concernée. Dans le cas contraire, la constitution de l'échantillon d'enquête visera à assurer une représentativité structurelle (on sur-représente a priori des individus présentant des caractéristiques rares et supposées intéressantes vis-à-vis de l'objet d'étude) ou une représentativité statistique (on cherche à rendre compte "exactement" de la population de départ).

L'enquête sur les prix du camionnage visait une représentativité structurelle, les expéditions du le bas de la hiérarchie devant y être sur-représentée.

Dans le cas d'une représentativité structurelle, on pourra constituer l'échantillon par la méthode des quotas. Elle consiste à se donner un certain nombre d'items (par exemple l'âge pour un individu, le secteur d'activité pour une entreprise) et, pour chaque modalité de chaque item, à fixer un nombre d'enquêtes à réaliser. Si l'on recherche une représentativité statistique, il conviendra de partir d'une base de sondage regroupant l'ensemble de la population étudiée et de générer un échantillon aléatoire simple. Un échantillon sera dit aléatoire si tous les individus de la population ont la même probabilité d'être sélectionnés, simple si les individus sont tirés indépendamment les uns des autres. La première méthode se révèle tout à fait suffisante dès lors que l'on se contente de valeurs relatives tandis que la seconde est la plus adaptée pour produire des informations en valeur absolue (et donc a fortiori en valeur relative). Mais elle impose de disposer d'une base de sondage exhaustive, ce qui n'est pas toujours le cas, et revient très cher.

2 - PRODUIRE LES INFORMATIONS

La production des informations sera abordée ici rapidement, à travers un certain nombre de remarques à propos de la collecte sur le terrain et des différentes étapes de la mise sur informatique.

2.1 - LA COLLECTE SUR LE TERRAIN

Le choix du personnel d'enquête est déterminant pour la qualité des résultats. Ce personnel ne se limite pas aux seuls enquêteurs. Leur activité doit être en effet contrôlée par un personnel spécialisé. Ces deux catégories de personnel doivent subir une formation théorique (présentation en chambre des objectifs de l'enquête, de son organisation, explicitation des différentes questions, ...) et une formation pratique (passation sur le terrain du ou des questionnaires avant le début réel de la phase d'enquête) qui peut conduire à en rejeter certains.

Ce test en situation des enquêteurs doit avoir été précédé de tests des questionnaires. La compréhension des objectifs de l'enquête et surtout des différentes questions doit en effet être évaluée préliminairement auprès d'un public similaire à celui qui subira ensuite l'enquête en vraie grandeur. Ce test est la dernière occasion pour des modifications de formulation, l'élimination ou le remplacement de questions. C'est également l'occasion de revenir sur la langue à utiliser, afin de trancher entre langue véhiculaire et langue vernaculaire.

2.2 - LA MISE SUR INFORMATIQUE

Le transfert des données, des bordereaux d'enquête aux fichiers informatiques de travail, se décompose en trois phases.

LE CODAGE

En matière de codage, une seule philosophie est à adopter : rechercher toujours la solution la plus simple pour le personnel de codage, l'ordinateur pouvant ensuite "travailler" pour recoder, regrouper, calculer, ... (voir plus loin).

Cette "philosophie" montre l'intérêt d'un précodage pour chaque variable l'autorisant. Pour les variables quantitatives, la codification doit bien évidemment correspondre à la valeur enregistrée, à un éventuel facteur d'échelle près. Pour les caractères qualitatifs, la codification peut être numérique (à chaque modalité correspond un et un seul nombre) ou alphanumérique (à chaque modalité correspond un et un seul intitulé, généralement de quatre caractères) : la première est moins coûteuse, la seconde plus "parlante".

> ***Les trois modalités du type de véhicule, caractère qualitatif, peuvent être codées par exemple UTLE, PORT et SEMI (codification alphanumérique) ou 1, 2 et 3 (codification numérique).***

Sur le plan matériel, le codage nécessite une équipe de codeurs, préalablement formés et testés (par exemple en codant les enquêtes issues du test des enquêteurs !). Un manuel de codage exhaustif est également indispensable, surtout lorsque l'équipe est importante, afin d'éviter les flottements et d'uniformiser au maximum les comportements. Le codage se traduit, sauf lorsqu'il a lieu simultanément à la saisie, par la création d'un bordereau de codage standardisé. Dans le cas d'enquêtes hiérarchisées, il convient bien sûr de prévoir plusieurs bordereaux et les manuels correspondants et surtout de prévoir des variables assurant la liaison entre les différents niveaux (voir plus haut).

LA SAISIE

La saisie permet de créer le ou les fichiers informatiques de base issu(s) de l'enquête.

La saisie de masse est la saisie "bête et méchante", au kilomètre. Elle implique, ou devrait impliquer, systématiquement une double saisie de chaque bordereau de codage. Elle s'oppose à la saisie contrôlée, qui permet de vérifier au fur et à mesure la validité des données. C'est alors simultanément une phase d'apurement : l'arbitrage est donc à faire entre le nombre de contrôles lors de la saisie, coûteux en temps et ralentissant celle-ci, et ceux lors de l'apurement proprement dit.

Là encore, un personnel spécifique doit être recruté, formé et testé (rapidité, fiabilité, ...).

LA PREPARATION DES DONNEES

La préparation des données peut être décomposée en deux phases : l'apurement, la génération de nouvelles variables.

L'apurement recouvre deux aspects: l'apurement aux bornes et l'apurement de cohérence.

Les caractères qualitatifs, de même que les variables quantitatives discrètes, n'admettent qu'un certain nombre de valeurs (numériques ou alphanumériques). Il importe alors de vérifier que les valeurs enregistrées sont bien des valeurs autorisées. On parle alors d'apurement aux bornes.

> ***Les valeurs licites pour le type de véhicule, si sa codification est numérique (voir supra), seront 1, 2 et 3. Toute autre valeur devra alors être signalée.***

L'apurement de cohérence a pour but de tester la cohérence entre les différentes informations recueillies lors de l'enquête. Ces croisements entre informations peuvent toucher des niveaux hiérarchiques différents.

> ***Si le véhicule utilisé par le transporteur pour une des expéditions recensées (information enregistrée au niveau expédition) n'apparaît pas dans le parc déclaré (information appartenant au niveau d'observation entreprise), une erreur doit être signalée.***

Attention, toutes les erreurs repérées à l'occasion de l'apurement (aux bornes ou de cohérence) ne sont pas nécessairement corrigibles. En effet, elles peuvent résulter d'un problème lié à la saisie mais aussi relever de la phase d'enquête elle-même, soit faute de l'enquêteur, soit cas particulier mal identifié lors de la rédaction du questionnaire et conduisant à des incohérences. Dans le premier cas, il suffit en fait de retourner vérifier les informations sur les fiches d'enquête et de comger les fichiers informatiques, mais dans le second cas, il n'y a le plus souvent pas d'issue satisfaisante.

La création des nouvelles variables est tout à la fois à la dernière étape de la préparation des données et la première étape de l'analyse. Certains indicateurs ne peuvent être recueillis lors de l'enquête mais peuvent être calculés à partir de données élémentaires. C'est par exemple le cas d'un prix à la tonne-kilomètre, rarement connu du transporteur, alors que le prix total perçu, le tonnage et la distance parcourue sont des informations mieux maîtrisées par l'opérateur. Il y a bien là au sens strict une préparation des données. Par contre, si l'on cherche à élaborer un indicateur de l'organisation spatiale des déplacements d'un individu (succession d'allers et retours au domicile, ou bien présence de déplacements secondaires, ...), on aura là au contraire les prémisses de l'analyse du comportement de mobilité des individus. La création de nouvelles variables est donc une phase extrêmement importante (et extrêmement longue aussi parfois !) qui montre bien, s'il en était encore besoin, que ces données ne sont guère données mais bien toujours à produire.

> ***Le prix à la tonne-kilomètre sera calculé automatiquement en divisant le prix total par le produit de la distance par la quantité.***

Le personnel nécessaire dans cette ultime phase de la production des données est sensiblement plus qualifié que pour les phases précédentes. Informaticiens ou statisticiens doivent être associés avec des chargés d'études ayant participé à la conception du questionnaire.

B - LES SYSTEMES D'INFORMATION STATISTIQUE

1 - DEFINITION ET OBJECTIFS

1.1 - DEFINITION

On peut définir un Système d'Information Statistique (SIS) comme la mise en oeuvre de moyens humains et matériels dans le but de produire régulièrement ou à la demande des données selon des procédures déterminées. Ces données sont stockées dans une Base (ou Banque) de données.

Moyens humains : en fonction de l'importance du système, la spécialisation du personnel sera plus ou moins poussée ; statisticiens spécialisés par secteur, informaticiens, opérateurs de saisie, ...

Moyens matériels : la baisse du prix des micro-ordinateurs et l'augmentation de leurs performances ont entraîné la généralisation de leur utilisation en matière de statistiques. On distingue :

- le matériel. Il s'agit des micro-ordinateurs et de leurs périphériques.
- les logiciels. Quatre types de programmes sont nécessaires.
  * un Gestionnaire de Bases de Données Relationnelles permettant de saisir, stocker et réaliser un premier traitement des données ;
  * un tableur permettant de réaliser des calculs et des tableaux complexes ;
  * un grapheur permettant de représenter graphiquement les données (intégré à la plupart des tableurs) ;
  * un traitement de textes pour produire des rapports.

L'utilisation d'un logiciel statistique spécifique ne s'impose que pour les traitements les plus complexes (analyse statistique).

Produire régulièrement : il est essentiel que les données soient produites à intervalles réguliers afin de constituer des séries chronologiques.

Produire à la demande : compte tenu de la masse d'informations stockées, il est souvent possible de réaliser tel ou tel traitement pour répondre à une demande spécifique.

Procédures déterminées : afin d'assurer la cohérence des données et d'éviter toute rupture de série chronologique, les données doivent être produites selon des procédures déterminées et répétitives (collecte des données, codage et saisie, traitement, ...).

Base (ou Banque) de données : le terme de "Base", contrairement à celui de Banque, fait explicitement référence à l'utilisation de l'outil informatique ; un gestionnaire de Bases de Données Relationnelles (cf. supra) permet d'accéder facilement aux données stockées dans différents fichiers. Il permet en particulier d'exploiter simultanément plusieurs fichiers de la base (immatriculations d'une part et licences de transport d'autre part, par exemple).

1.2 - OBJECTIFS

La constitution d'un SIS renvoie à quatre objectifs.

L'Etat pour contrôler et prendre des décisions doit connaître avec précision le secteutr concerné. Les mots "Etat" et "statistique" ont d'ailleurs la même origine.

Un Système d'Information Statistique permet un suivi systématique du système de transport et d'en repérer les éventuelles transformations : prix de transport, coûts, parts du marché entre transporteurs nationaux et étrangers en transport international, ...

Il est également une source d'informations pour les études sur le secteur des transports. Diverses études peuvent utiliser les mêmes données de base. Dans un souci de cohérence, il est important qu'elles utilisent les mêmes sources.

Enfin, les données que produit un SIS peuvent être utiles dans d'autres domaines : compatibilité nationale, commerce extérieur, ...

2 - LES DIFFERENTES DONNEES A RECUEILLIR ET LEURS SOURCES

2.1 - LES SOURCES

Deux types de sources peuvent être utilisées, les enquêtes et les fichiers administratifs.

Les enquêtes, lorsqu'elles sont réalisées selon les règles de la statistique, sont plus fiables que les sources administratives mais peuvent se révéler extrêmement coûteuses.

La production de données par des enquêtes est particulièrement coûteuse. Pour ce type de données, il vaut mieux utiliser des sources administratives en les complétant, éventuellement, par des enquêtes ponctuelles.

Il est évident qu'il n'est pas du ressort d'un service statistique d'informatiser l'ensemble des fichiers administratifs. Ceux-ci doivent lui être transmis par les services concernés : Service des Cartes grises, des Licences de Transport, ... Une collaboration étroite entre le service statistique et les autres services est alors nécessaire. Il est souvent souhaitable de faire remplir un questionnaire statistique à l'occasion d'un acte administratif.

2.2 - LES DIFFERENTS TYPES DE DONNEES

Après une vue d'ensemble, nous présentons en détail les données relatives au transport routier non urbain.

VUE D'ENSEMBLE

Les données à produire dépendent des besoins et des coûts de leur production mais également de l'organisation spécifique du système de transport dans chaque pays. Un pays sans voie ferrée n'aura pas à se préoccuper de données ferroviaire, par exemple.

Un SIS peut être amené à couvrir les domaines suivants :

- transport aérien : flux de passagers, fret et poste par origine destination, mouvements d'aéronefs, parcs d'aéronefs, ...
- transport maritime : trafics par port d'embarquement, par compagnie maritime, taux de fret, ...
- transport fluvial et lacustre : nombre et capacité des péniches ou pinasses, trafic origine-destination, ...
- transport ferroviaire : nombre et caractéristiques des wagons, caractéristiques des voies, ...
- transport urbain : parc de transport public, trafic, accidents, coûts de transports, tarifs, ...
- Transport routier non urbain (voir infra).

Le nombre de données qui peuvent avoir à être centralisées dans une Banque de Données est donc considérable. Nous nous focaliserons dans la suite de l'exposé sur le domaine des transports routiers non urbains.

LES DONNEES RELATIVES AU TRANSPORT ROUTIER NON URBAIN

Les données relatives aux transports aériens, maritimes et ferroviaires sont plus faciles à produire et plus fiables que les données relatives au transport routier car il s'agit d'un transport beaucoup plus concentré : nombre limité de points de chargement/déchargement, grandes entreprises, ...

En ce qui concerne le transport routier non urbain, cinq grandes catégories de données peuvent être recensées.

Offre de transport

Les données relatives à l'offre concernent les infrastructures routières, les entreprises du secteur, le parc de véhicules et leurs coûts d'exploitation.

* Infrastructures routières

Source : inventaire routier

Données à recueillir par tronçon :

- kilométrage
- largeur
- nature de la chaussée
- état de la chaussée
- pente et courbure.

* Entreprises de transport

Source : fichier des licences ou des autorisations de transport

Données à recueillir :

- nombre et taille des entreprises
- créations - disparitions d'entreprises (difficile)
- activités.

* Parc de véhicules

Sources : fichier des immatriculations, des licences de transport, des visites techniques, des vignettes, ...
Données à recueillir :

immatriculations et ré-immatriculations
données sur le parc en fonction des types de véhicules, de leur capacité (passager ou charge utile), de leur marque, de leur âge.

* Coûts d'exploitation des véhicules (données nécessaires à leur élaboration)

Sources : enquêtes, lettres de Voitures Obligatoires
Données à recueillir :

- données physiques : kilométrages, consommation de pneus, de pièces détachées, de carburant, ...
- données monétaires : prix du carburant, des pneus, ...

Demande de transport

Les données concernent ici les trafics routiers et les flux de transport.

* Trafics routiers

Source : enquêtes de trafic
Données à recueillir :

Trafics par types de véhicules pour chaque tronçon du réseau.

* Flux de transport

Sources : enquêtes origine-destination (transport national), documents douaniers (transport international), lettres de Voiture Obligatoires (national et international)
Données à recueillir :

- Enquêtes O-D : origine-destination des véhicules et des marchandises (en fonction de leur nature et de leur poids). Le temps de transport est également une donnée nécessaire mais difficile à recueillir.
- Documents douaniers : flux origine-destination par marchandise (nature, poids, valeur), nationalité des véhicules.
-Lettres de Voiture Obligatoires : toutes ces données en transport national comme international.

Prix de transport

Il s'agit de données nécessaires à la mesure des performances du système de transport.

Sources : lettres de voiture, enquêtes
Données à recueillir : prix à la tonne.km, par marchandise, véhicule, origine-destination.

Intermédiaires de transport

* Courtier de fret

Source : licences de courtiers de fret, enquêtes
Données à recueillir : nombre de courtiers par centre et domaine d'activité.

* Transitaires

Source : autorisations en douane, enquêtes
Données à recueillir : nombre, activités, chiffre d'affaires.

Accidents de la circulation en rase campagne

Sources : constats réalisés par la police ou la gendarmerie, assurances.
Données à recueillir : nombre d'accidents par types, nombre d'accidentés, véhicules impliqués, condition de circulation. En matière d'accidents, les données à recueillir sont particulièrement nombreuses.

3 - UNITES DE MESURE ET CLASSIFICATIONS.

3.1 - UNITES DE MESURE

Certaines unités de mesure sont couramment employées en matière de statistiques sur les transports routiers. L'utilisation de ces unités pose néanmoins certains problèmes.

MESURE DE LA CAPACITE DU PARC EN TONNES DE CHARGE UTILE

La charge utile d'un véhicule dépend :
- des caractéristiques techniques des véhicules indiquées par le constructeur,
- de la réglementation sur le poids des véhicules.

La capacité de charge utile d'un parc est alors traditionnellement égale à l'addition de la charge utile de tous les véhicules qui le composent.

Mais compte tenu des surcharges pratiquées en Afrique (de 50 à 100 %), on sous-estime ainsi la capacité réelle de charge utile ce qui fausse la comparaison offre-demande de transport.

Si l'on prend la charge effective maximum des véhicules, on risque par contre de surestimer cette capacité (rapport charge utile/volume utile trop important pour de nombreuses marchandises).

MESURE DES DISTANCES EN KM

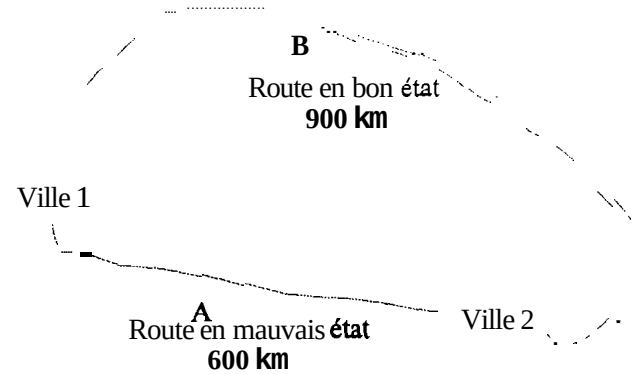
La mesure des kilomètres peut être a priori réalisée de quatre manières différentes :

- à vol d'oiseau,

- en fonction des kilomètres réellement parcourus par les véhicules,
- en fonction du trajet le plus court,
- en fonction du chemin optimal tel que l'on peut le calculer en fonction des distances et des coûts d'exploitation des véhicules sur chaque tronçon du réseau.

Si l'on prend la distance réellement parcourue par les véhicules, on peut sous-estimer certains prix de transport mesurés en tonne.km.

Mais si l'on choisit d'affecter un kilométrage à vol d'oiseau, par le trajet le plus court ou par le trajet optimal, les données produites ne seront plus cohérentes avec celles sur le kilométrage annuel des véhicules.



Un transport de 30 tonnes est payé 1 000 000 FCFA, quel que soit le trajet choisi par le transporteur.
S'il choisit le trajet A, le prix de transport sera de 55 FCFA la tonne.km, s'il choisit le trajet B, il sera de 41 FCFA la tonne.km.

MESURE DES FLUX DE TRANSPORT EN TONNE.KM

Une tonne.km est égale, par définition, à une tonne de marchandises transportée sur un kilomètre.

Elle combine donc deux unités :

* la tonne

Il est possible de mesurer les flux en tonnes mais cela entraîne

1) une sous-estimation des transports longue distance,
2) une sur-estimation du transport des produits pondéreux.

* le kilomètre

Il est également possible de mesurer les flux en nombre de kilomètres réalisés par les véhicules mais cela entraîne

3) une sur-estimation des expéditions de faible tonnage,
4) une sur-estimation des transports longue distance.

L'utilisation de la tonne.km permet de résoudre les points 1 et 3 mais pas les points 2 et 4. Elle sur-estime les transports de charges importantes sur de longs trajet. Or, un transport par camion de 30 tonnes de coton sur 1 000 km n'est pas équivalent, du point de vue du service produit, à 100 transports par pick-up, de 3 tonnes, de produits maraîchers sur 100 km. Les différences de prix constatées aux différents niveaux de la hiérarchie du système de transport sont liées à l'utilisation de la tonne.km comme unité de mesure, qui considère comme équivalents des services pourtant très

différents.

Derrière l'apparente neutralité de unités de mesure se cachent donc des choix, bien souvent implicites. L'usage des unités usuelles reste bien évidemment possible et même indispensable, il suffit de garder à l'esprit les conséquences de ces choix.

3.2 - CLASSIFICATIONS

On distingue deux types de classification :
- les classifications déductives ; on part des données brutes pour les regrouper en un certain nombre de classes (voir supra) ;
- les classifications inductives ; on commence par établir une classification avant de produire les données.

L'adoption d'un certain nombre de classifications inductives est nécessaire au fonctionnement d'un SIS.

APPLICATION DES CLASSIFICATIONS HABITUELLES EN STATISTIQUES TRANSPORT AU PAYS AFRICAINS.

Certaines classifications habituellement employées en statistique transport ne sont pas entièrement satisfaisantes lorsqu'on les applique aux systèmes de transport des pays africains.

* Transport de personnes, transport de marchandise.

L'importance du transport mixte, en particulier dans les pays sahéliens, rend souvent cette classification non pertinente.

* Transport pour un compte propre, transport pour compte d'autrui

De nombreuses entreprises sont à la fois des entreprises commerciales transportant des produits qui leur appartiennent (transport pour compte propre) et des entreprises de transport proprement dites (transport pour compte d'autrui). Dans ces conditions, il est difficile de déterminer précisément le nombre d'entreprises de transport. Il est également difficile de déterminer l'effectif du parc de transport pour compte d'autrui.

LA REALISATION DES NOMENCLATURES

L'établissement d'une nomenclature est nécessaire lorsque l'on doit réaliser une classification complexe et qu'il faut tenir compte d'un très grand nombre de cas différents.

Une nomenclature est organisée selon une structure arborescente permettant de réaliser des regroupements de postes plus ou moins importants. Un code est ensuite affecté à chaque poste, code permettant de repérer simultanément le niveau du poste et sa nature.

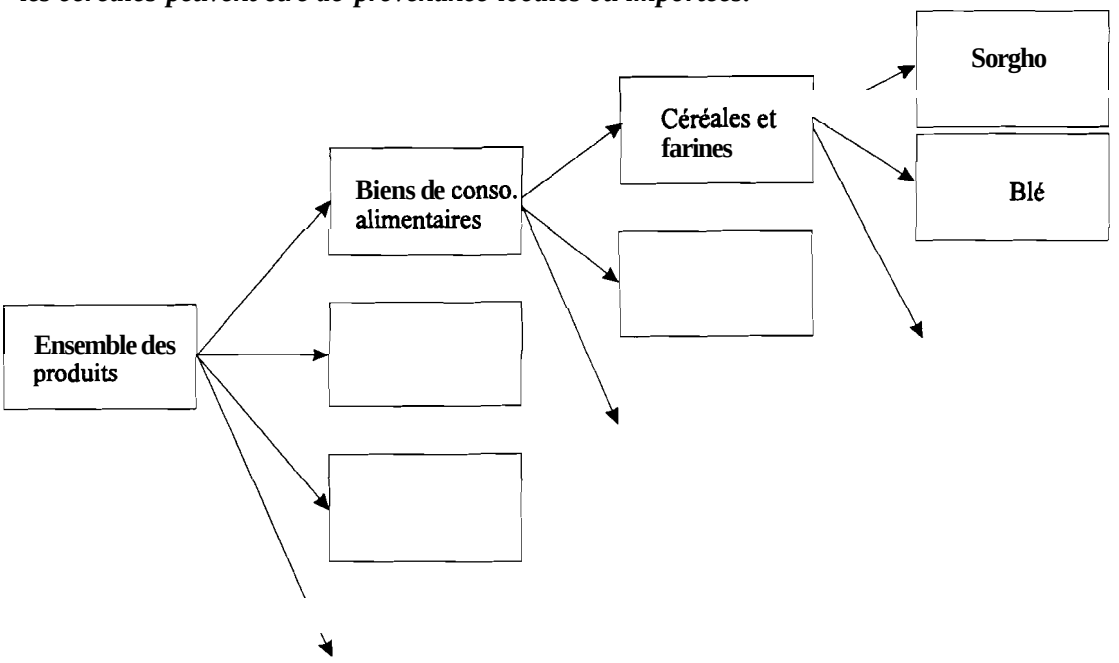
Nous prendrons ici l'exemple de la classification des produits transportés.

Dans toute la mesure du possible, une nomenclature doit répondre à deux conditions :

- équilibre entre les postes,
- homogénéité.

En ce qui concerne une Nomenclature des Produits Transportés, il faut que chaque poste regroupe des marchandises transportées dans des conditions homogènes. On peut par exemple regrouper le coton et les engrais. Mais ces conditions ne peuvent pas toujours être respectées.

* ***l'arachide peut être*** à ***la fois un produit vivrier et un produit agro-industriel.***
* ***les céréales peuvent être de provenance locales ou importées.***



01 biens de consommation alimentaire
01:01 céréales et farines
01:01:01 ***sorgho.***

Une nomenclature doit tenir compte de la spécificité de chaque pays. Elle n'est jamais totalement adaptée à un objet d'étude précis.

4 - FIABILITE DES SOURCES ADMINISTRATIVES

Nous avons déjà souligné que les données produites par enquête était plus fiables que celles issues de documents administratifs mais qu'elles étaient très coûteuses si l'on voulait produire des données en valeur absolue (enquête origine-destination par exemple).

Les sources administratives sont normalement des sources exhaustives mais le non-respect de la réglementation se répercute sur la fiabilité des données.

- ***Transporteurs travaillant sans licence,***
- ***Sous-estimation des prix indiqués sur les LVO,***
- ***Sur-estimation du parc (non retrait des cartes grises).***

Il est par contre souvent possible de redresser les données issues des sources administratives en comparant les données ou en réalisant des enquêtes complémentaires.

C - LES METHODES STATISTIQUES.

Cette troisième partie comportera 5 sections. La première commence par un petit rappel historique dont l'objectif consiste à démystifier quelque peu une discipline scientifique vaste dont nous n'en étudierons qu'une partie limitée : "la statistique descriptive". C'est pourquoi nous préciserons dans un deuxième point les limites de ce cours de méthode. Enfin, il est essentiel d'apprendre à développer un sens critique par rapport à l'emploi de l'outil statistique. C'est pourquoi nous avons repris, dans un troisième point, une classification des erreurs statistiques les plus courantes proposée par J. KLATZMANN, dans son ouvrage "Attention statistiques !"([1]).

La deuxième section nous permettra de préciser un certain nombre de définitions essentielles. Il sera également l'occasion de passer en revue les principaux outils de traitement statistique uni-varié. Ces outils permettent l'analyse d'une population statistique au regard d'un seul caractère. On distinguera tout d'abord les tableaux et les graphiques, puis nous présenterons une série d'indicateurs de tendance centrale et d'indicateurs de dispersion.

La troisième section traitera de la statistique bi-variée. Nous apprendrons cette fois à utiliser les instruments statistiques qui permettent l'observation d'une population au regard de deux caractères pris simultanément.

Les deux dernières sections aborderont les méthodes factorielles et de classification.

Chacune de ces sections sera illustrée d'exemples extrait de l'enquête sur la mobilité réalisée par le LET à Ouagadougou en février 1992. L'échantillon d'enquête porte sur 3682 individus de14 ans et plus.

1. QUELQUES PRECISIONS IMPORTANTES

1.1. UN PETIT HISTORIQUE DE LA STATISTIQUE

La statistique a d'abord été descriptive et a consisté à l'origine à observer des faits. Les premiers recueils de données remontent à l'antiquité. Ce sont les chefs d'Etat, qui les premiers éprouvèrent le besoin de dénombrer les éléments de leur puissance : population, potentiel militaire, richesses... Les premiers recensements connus sont apparus avec la civilisation sumérienne, de 5000 à 2000 ans avant notre ère. On possède, en effet, des listes d'hommes et de biens inscrits sur des tablettes d'argile. De même, le relevé des personnes et des biens a lieu régulièrement en Mésopotamie 3000 ans avant Jésus-Christ. Enfin, l'Egypte semble avoir été la première nation à organiser des recensements systématiques de population au moins depuis l'an 2900 avant J-C, mais également à institutionnaliser des recensements à finalité fiscale (2700 à 2500 avant J-C). Sous le Pharaon AMASIS II, au VI^{ème} siècle avant notre ère, tout individu était tenu de déclarer ses sources de revenu et son activité. Tout manquement à cette règle était puni de mort.

---

[1] J. KLATZMANN, "Attention statistiques ! Comment déjouer les pièges", Paris, Editions La Découverte, essais, 1992, 251 p.

Si la statistique est très ancienne, l'apparition du mot statistique lui-même, est relativement récent, même s'il est très difficile de l'attribuer à quelqu'un en particulier. Il existe une Biblioteca Statistica datant de 1701 et un Microscopium Statisticum remontant à 1672. On peut encore trouver des traces du mot statistique dans le langage administratif français colbertien à travers une Déclaration des biens, charges, dettes et statistiques des communautés de la généralité de Bourgogne, établie par un certain Claude BOUCHU, intendant de Bourgogne de 1666 à 1669.

1.2. STATISTIQUE DESCRIPTIVE OU STATISTIQUE INDUCTIVE

LA STATISTIQUE DESCRIPTIVE

La statistique descriptive résume, récapitule, analyse un ensemble de données à l'aide d'un ou deux éléments d'information qui caractérisent la totalité de ces données. On distingue, tout d'abord, la statistique uni-variée qui permet d'observer une population au regard d'un caractère et la statistique bi-variée qui consiste à étudier une population en croisant deux caractères. Il est également possible d'étudier une population en considérant simultanément toute une série de critères. On sort alors du domaine de la statistique purement descriptive pour entrer dans ce que l'on appelle communément "l'analyse de données". Cette approche propose une panoplie d'instruments assez riche et recourt régulièrement aux représentations graphiques et à la géométrie. L'expansion de l'informatique a largement contribué au développement de ces techniques qui remportent aujourd'hui un large succès.

LA STATISTIQUE INDUCTIVE

La statistique inductive s'attache à deux catégories de problèmes : les problèmes d'estimation visant à généraliser à une population les résultats obtenus sur un échantillon extrait de cette population, les tests d'hypothèse ayant pour but d'accepter ou de rejeter la validité d'une hypothèse. Quelle que soit la méthode, les échantillons doivent être aléatoires et simples. De plus, des hypothèses sur la distribution de probabilité du phénomène dans la population mère sont souvent nécessaires.

Exemples de problèmes : construction d'un intervalle de confiance pour estimer une moyenne (problème d'estimation) ou test de conformité d'une moyenne (test d'hypothèse), tests d'ajustement d'une distribution empirique à une distribution théorique, test d'indépendance de deux caractères qualitatifs, test de l'égalité des moyennes entre elles dans le cadre d'une analyse de la variance, utilisation de la régression linéaire à des fins de prévision, ...

En dépit des hypothèses lourdes portant sur la nature des distributions dans les populations mères (l'hypothèse de normalité par exemple), ces techniques demeurent encore largement utilisées, le meilleur exemple étant le développement de l'économétrie.

1.3. SOURCES D'ERREURS COURANTES EN STATISTIQUE

Joseph KLATZMANN, ancien administrateur à l'INSEE, insiste sur les 11 pièges les plus fréquents tendus par cette discipline.

LES STATISTIQUES INEVITABLEMENT FAUSSES

Dans le cas d'une enquête, il n'est pas rare que les personnes interrogées n'aient pas les moyens (questions mal posées) ou n'aient pas le désir de répondre (questions trop personnelles). De même, est-il possible qu'une personne n'ayant pas d'opinion sur une question se sente obligée de répondre. Quelle valeur aura cette réponse ?

A la lecture des résultats d'enquêtes, il convient également de prendre un certain nombre de précautions. La population de la France peut-elle être connue avec exactitude ?

LES STATISTIOUES TRUQUEES

Le producteur de statistiques peut avoir dans certains cas intérêt à publier des résultats faux, ou plus exactement gonflés ou baissés artificiellement. Le meilleur exemple est constitué par les chiffres du chômage qui donnent lieu à de savantes manipulations de la part des gouvernements, notamment à l'approche de seuils psychologiques. Les manipulations reposent dans ces cas là sur une modification de certaines définitions.

LES STATISTIQUES DONT ON TIRE PLUS QU'ELLES NE PEUVENT DONNER

Le fait d'interroger un millier de personnes au cours d'un sondage, permet de ce faire une idée de l'opinion d'une population beaucoup plus importante. Mais l'interprétation qui vaut pour une population importante ne peut être interprétée de la même façon pour des groupes plus restreints. Que dire de l'opinion des agriculteurs lorsque sur 1000 personnes interrogées, sept sont agriculteurs. Compte tenu de leur place actuelle dans la population française, Joseph Klatzmann souligne qu'il faudrait interroger plus de 15000 personnes pour trouver parmi elles 1000 agriculteurs.

UES "VRAIES-FAUSSES"

Il s'agit de statistiques exactes qui présentées d'une certaine manière peuvent donner lieu à des interprétations erronées. Pour pouvoir réaliser un graphique plus lisible, on peut souhaiter que les abscisses ou les ordonnées ne commencent pas à zéro. Cependant les variations observables de la courbe représentée peuvent donner l'illusion d'être plus importantes qu'elles ne le sont en réalité.

LES AFFIRMATIONS SANS AUCUN FONDEMENT

Attention aux affirmations intempestives. Il suffit qu'un chiffre ait été repris plusieurs fois pour qu'il devienne quasi-officiel. Personne dès lors ne le remet plus en doute. Joseph Klatzmann prend l'exemple des 50 millions de morts de faim par an qui, il y a quelques années, a frappé les imaginations. A l'époque, le nombre total de décès par an, toutes catégories confondues, était inférieur à ce chiffre.

LA PRECISION ILLUSOIRE

Beaucoup de résultats ne sont que des ordres de grandeur. Cela n'empêche pas de nombreuses personnes de publier ces résultats avec un maximum de "précision". Le cas le plus fréquent concerne les pourcentages fournis avec un, voire deux chiffres après la virgule. Le type d'enquêtes permettant d'obtenir certains résultats et les risques d'erreurs importants devraient nous inciter à une plus grande prudence.

LES MOYENNES SANS SIGNIFICATION

Le recours aux moyennes permet de gagner en lisibilité. Il convient toutefois d'être prudent. Quel intérêt constitue la comparaison de la densité de la population en Egypte et en France, dans la mesure où en Egypte la quasi-totalité du territoire est un désert. La densité peut ainsi paraître faible alors qu'au contraire, elle est importante dans les zones habitables.

LES CALCULS AUX RESULTATS IMPRESSIONNANTS, MAIS QUI NE VEULENT RIEN DIRE

Beaucoup de gens sont friands de slogans du type :"il y a un chômeur de plus toutes les deux minutes". Ce genre d'expression n'est pas sans comporter certains risques. En effet, est-on capable d'évaluer ce que cela représente ? De quel pays s'agit-il ? Quelle est sa population active ? Combien d'emplois ont été créés dans le même temps ?... Ce genre de simplification est pour le moins dangereux.

LES COMPARAISONS DE CHOSES NON COMPARABLES

L'exemple le plus classique consiste à comparer les revenus entre deux pays. Dans l'un on peut vivre raisonnablement avec une certaine somme. Dans l'autre cette somme paraît dérisoire.

LES FAUSSES CORREJ.ATIONS

Une corrélation est une relation positive ou négative entre deux phénomènes. Cette relation n'est pas absolue. La corrélation exprime souvent une relation de cause à effet. Si dans certains cas cette relation est évidente, dans d'autre cas elle est bien difficile à affirmer. Parfois chacun des deux phénomènes est à la fois la cause et l'effet, mais parfois la corrélation ne résulte d'aucune cause ou effet entre les deux phénomènes observés. L'interaction d'un troisième phénomène peut alors expliquer les variations observées.

LES PROBLEMES COMPLEXES, LES "SUBTILITES" DE LA STATISTIQUE

Il existe enfin un certain nombre de pièges dans lesquels le statisticien n'est jamais à l'abri de tomber : "Dans chaque région, on constate que les agriculteurs consomment par personne plus de mil que les non agriculteurs. En revanche, au plan national, la consommation des non-agriculteurs est supérieure à celle des agriculteurs". Il suffit pour cela que les agriculteurs soient concentrés dans les régions où la consommation de mil est la plus faible.

Il est essentiel de bien faire la différence entre "la statistique" et "les statistiques". La statistique regroupe l'ensemble des méthodes scientifiques qui permettent d'analyser quantitativement l'information. Elle est de ce point de vue parfaitement rigoureuse. Les statistiques

ne sont, en revanche, que les résultats numériques auxquels conduit l'application de ces méthodes. Les chiffres peuvent être faux, mal recueillis ou mal interprétés (volontairement ou non). Les méthodes et processus techniques ne sont ni faux, ni truqués. Tout au plus pourront-ils être mal choisis face à un objectif donné.

2. STATISTIQUE DESCRIPTIVE UNI-VARIEE

La tâche principale du statisticien consiste à arbitrer entre la recherche d'une plus grande lisibilité et le maintien d'une certaine richesse d'information. Il est souvent difficile pour ne pas dire impossible de tirer des informations d'un tableau brut constitué, par exemple, à partir des résultats d'une enquête : imaginons la lecture d'un tableau concernant une centaine d'individus ayant répondu à une trentaine de questions !!! Le premier travail du statisticien va consister à faire des choix, par rapport à ses propres objectifs, ses propres hypothèses, pour extraire une information et la rendre plus lisible. Il dispose pour ce faire de toute une panoplie d'outils. Les tableaux et les graphiques sont les premiers qui nous viennent à l'esprit. Ils constituent pour le statisticien, deux types privilégiés de lecture des résultats. Si les tableaux présentent l'avantage de la simplicité, les graphiques offrent une forme imagée souvent très utile.

2.1. QUELQUES DEFINITIONS FONDAMENTALES

Nous avons vu que l'une des premières disciplines à avoir recouru aux statistiques était la démographie. Il n'est donc pas surprenant de rencontrer des termes tels que la population ou les individus pour décrire les ensembles sur lesquels nous allons travailler. Il conviendra toutefois de ne pas restreindre l'intervention de la statistique au domaine démographique. Population et individus doivent être pris au sens large du terme.

POPULATION STATISTIQUE

Les ensembles étudiés de manière quantitative portent le terme générique de population. Lorsque l'on parle de population étudiée, il peut s'agir de la population française repérée par le dernier Recensement, mais il peut également s'agir de marchandises expédiées par une entreprise de messagerie, de véhicules composant la flotte d'un transporteur particulier, d'accidents de la route qui se sont produits au cours d'une certaine période, de pays européens,...

Suivant les cas, le statisticien se trouvera confronté à un ensemble d'êtres humains, à des stocks ou des flux d'objets concrets, à des ensembles de biens immatériels ou encore à des ensembles non-concrets. Chaque population contiendra ainsi un ensemble fini d'éléments pouvant être dénombrés. Enfin une population doit être définie de manière très précise. Les frontières de l'ensemble soumis à l'analyse statistique doivent en particulier être délimitées avec précision.

INDIVIDU STATISTIQUE

Chaque observation réalisée par le statisticien sur la population porte sur un individu statistique. Les éléments composant la population sont de même nature, mais peuvent être fort différents d'une population à l'autre. Il peut s'agir de chômeurs, de jours d'une année, de déplacements d'individus ou de marchandises...

LA CARACTERISATION D'UN INDIVIDU

Un élément composant une population peut être caractérisé de très nombreuses façons. Un individu peut être masculin ou féminin, être d'un certain âge, avoir un certain nombre d'enfants, disposer d'un certain revenu... Chacune de ces caractéristiques présente des spécificités et nécessite un traitement propre. Mais la plupart du temps, pour pouvoir décrire de façon efficace un ensemble d'individus on va procéder à des regroupements en sous-ensembles appelés caractères.

Les caractères

Pour décrire quantitativement une population, on s'efforce de classer les individus qui la composent en sous-ensembles. Ce classement peut se faire relativement à un ou plusieurs caractères. Le statisticien aura pour tâche de ne retenir que les caractères les plus pertinents, en agrégeant les informations en sous-ensembles cohérents.

Les modalités

Chaque caractère étudié peut présenter deux ou plusieurs situations différentes que l'on appelle modalités. Pour que le classement d'une unité statistique soit toujours possible sans ambiguïté, les différentes modalités doivent vérifier les critères d'exhaustivité et d'incompatibilité. L'incompatibilité signifie qu'aucun individu ne peut entrer dans deux modalités d'une variable. L'exhaustivité signifie que tout individu doit pouvoir être classé dans une des modalités proposées. Le respect de ces deux principes garantit le fait que la somme des effectifs de toutes les modalités d'une variable donne l'effectif total de la population. Pour la même raison, il garantit que la somme des fréquences relatives est égale à 1 (ou 100%). Tout individu ne peut être compté qu'une seule fois (incompatibilité) et tout individu doit avoir été compté au moins une fois (exhaustivité).

Il arrive que l'on ait pas de renseignements sur un ou plusieurs individus : il suffit qu'ils n'aient pas répondu à la question posée ou que l'on ne soit pas parvenu à se procurer l'information. Deux solutions sont envisageables lors du traitement. La première consiste à créer une modalité supplémentaire à la variable. Suivant les cas les individus seront rangés dans la catégorie : non réponse, non valide, ne sait pas, sans réponse..; Dans la seconde, on se limite à l'ensemble des individus ayant répondu. Dans ce cas, au lieu de travailler sur la population de départ, on travaillera sur la sous-population des individus ayant fourni une réponse.

Enfin il sera possible de hiérarchiser les modalités selon le degré de finesse de l'information disponible ou recherchée.

Caractere auantitatif - caractère aualitatif

Un caractère peut être soit qualitatif, soit quantitatif. Dans ce dernier cas, on lui associe une variable statistique. Cette différence est fondamentale dans la mesure où elle donne lieu à des traitements distincts de l'information. Le calcul d'indicateurs de tendance centrale ou de dispersion n'auront, en particulier, de sens que pour les variables quantitatives.

Des **caractères** sont **quantitatifs** lorsque leurs modalités sont mesurables, c'est-à-dire lorsque ces modalités peuvent être traduites par des nombres qui en mesurent les valeurs. Le caractère quantitatif prend alors le terme de variable statistique. Les modalités figurent alors les différentes valeurs possibles de la variable.

Les **caractères qualitatifs** sont des caractères dont les modalités échappent à la mesure. Celles-ci peuvent simplement être constatées.

Il est toujours possible de transformer une information de type quantitatif, en information de type qualitatif. L'inverse est beaucoup plus exceptionnel et correspond à des cas particuliers comme la création d'un tableau logique.

Variable quantitative discrète - variable quantitative continue

Une **variable** est dite **discrète** lorsqu'elle ne peut prendre qu'un nombre limité de valeurs à l'intérieur de son intervalle de variation. Les modalités du caractère sont soit des valeurs exactes, soit des regroupements de valeurs en classes. Le nombre d'enfants par ménage, le nombre de déplacements quotidiens, le nombre de colis expédiés ou reçus par une entreprise, le nombre d'employés d'une entreprise,..., sont des variables quantitatives discrètes.

A l'inverse, une **variable** est dite **continue** lorsqu'elle peut prendre un nombre illimité de valeurs à l'intérieur de son intervalle de variation. Une telle variable donne lieu, dans la réalité, à des regroupements en classes. Le poids ou la longueur d'un colis, le chiffre d'affaires, le salaire,..., sont des variables quantitatives continues.

La frontière entre le "discret" et le "continu" est illusoire et artificielle, compte tenu de la précision des techniques de mesure. Le regroupement des valeurs en classes donne lieu à deux problèmes particuliers :

- le choix de l'amplitude qui pourra être constante ou variable,
- la définition des extrémités de classes, en particulier la borne inférieure de la plus faible classe et la borne supérieure de la plus forte classe pourront être définies a ***priori*** ou ***a posteriori.***

2.2. L'ANALYSE D'UN CARACTERE QUALITATIF

LA CONSTRUCTION D'UN TABLEAU DE FREQUENCE

Nous allons réaliser une partition de la population de départ sur la base de l'étude d'une variable qualitative. Chaque individu statistique va se trouver rangé dans un sous-ensemble correspondant à l'une des modalités proposées. On va alors s'intéresser aux fréquences qui représentent la part de l'effectif correspondant à une modalité par rapport à l'effectif total. Dans la littérature statistique, on trouvera souvent la distinction entre fréquences absolues (correspondant aux effectifs de chaque modalité) et fréquences relatives (effectifs des modalités ramenés à l'effectif total de la population).

L'effectif total est : $N = \sum_i n_i$

La fréquence d'une classe i s'écrit : $f_i = n_i / N$

Exemple : la répartition par sexe

Parmi l'échantillon de l'enquête sur la mobilité à Ouagadougou, on va s'intéresser à la proportion d'hommes et de femmes. Le caractère étudié "sexe de l'individu" va donner lieu à deux modalités "masculin" et "féminin". Il est dès lors possible de construire un tableau présentant les

fréquences absolues et relatives. Nous avons ajouté une colonne libellant les fréquences relatives en pourcentages.

Modalités	Fréquences absolues n_i	Fréquences relatives f_i	Fréquences relatives (%) f_i*100
Féminin	1698	0.46	46
Masculin	1984	0.54	54
Total	3682	1	100

On peut aisément vérifier que :

- $N=\sum_i n_i=n_1+n_2=1698+1984=3682$
- $f_1=n_1/N=1698/3682=0.46$
- $f_2=n_2/N=1984/3682=0.54$

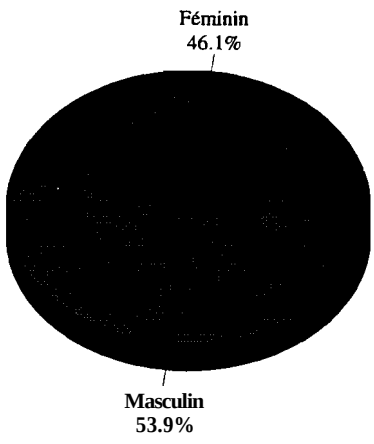
LE CHOIX D'UN GRAPHIOUE

Nous le verrons plusieurs graphiques peuvent être utilisés pour représenter la distribution de la population au regard d'un caractère qualitatif.

Le diaeramme circulaire

Le plus courant est le diagramme circulaire, vulgairement appelé camembert en raison de sa forme. Il s'agit d'un cercle découpé en parts (ou secteurs) dont les surfaces sont proportionnelles à l'effectif qu'elles représentent. Sachant que la circonférence d'un cercle est égale à $2\Pi R$, lorsque l'on multiplie cette longueur par les fréquences relatives libellées en pourcentage, on obtient la part de cercle dont chaque modalité doit disposer pour avoir une surface proportionnelle à son effectif.

Répartition par sexe

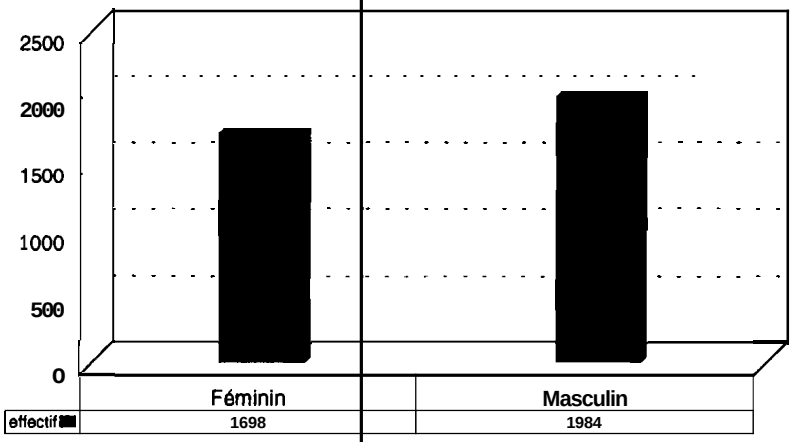


Source : enquête ménage Ouagadougou 92

Le diagramme en bâtons ou tuyaux d'orgue

Celui-ci est également assez courant. Dans ce cas, les modalités sont représentées par de simples bâtons, ou par des rectangles de base constante et de hauteur proportionnelle aux effectifs (ou aux fréquences).

Répart tion par sexe



Source : enquête ménage Ouagadougou 92

Les autres représentations

Parmi les nombreuses autres représentations possibles, on peut encore citer les diagrammes figuratifs qui consistent à présenter la distribution du caractère à l'aide d'illustrations. Cependant ce type de représentations présente certains pièges. Le lecteur ne sait pas si il doit prendre en considération les longueurs, les surfaces ou les volumes du diagramme figuratif. Ce genre de représentation aboutit souvent à une sur-représentation des grands effectifs, comme l'indique le graphique qui suit. Ceci se comprend facilement car on représente un phénomène à une dimension à l'aide d'une représentation à deux dimensions.

2.3. L'ANALYSE D'UN CARACTERE QUANTITATIF

Tout au long de cette présentation, nous allons retrouver la distinction évoquée entre variables quantitatives discrètes et variables quantitatives continues.

CARACTERE QUANTITATIF DISCRET

La détermination des fréauences relatives

On s'intéresse ici aux variables qui ne peuvent prendre qu'un nombre de valeurs limité à l'intérieur de leur intervalle de variation. Les individus statistiques sont rangés en sous-ensembles correspondant aux différentes valeurs de la variable X. Ces valeurs jouent le même rôle que les modalités des caractères qualitatifs. On pourra présenter les résultats relativement aux données brutes (fréquences absolues) ou en rapportant les effectifs correspondant à chaque valeur de X, à l'effectif total (fréquences relatives). Il sera, comme précédemment, possible de regrouper certaines valeurs de X.

Exemple la répartition des individus selon leur mobilité quotidienne :

Numéro de la classe i	Valeur de la variable X_i	Fréquences absolues n_i	Fréquences relatives f_i	Fréquences relatives (%) f_i*100
1	0	423	0.1149	11.49%
2	1	34	0.0092	0.92%
3	2	1024	0.2781	27.81%
4	3	109	0.0296	2.96%
5	4	1042	0.2830	28.30%
6	5	217	0.0589	5.89%
7	6	466	0.1266	12.66%
8	7	119	0.0323	3.23%
9	8	119	0.0323	3.23%
10	9	45	0.0122	1.22%
11	10	28	0.0076	0.76%
12	11	15	0.0041	0.41%
13	12	17	0.0046	0.46%
14	13	8	0.0022	0.22%
15	14	3	0.0008	0.08%
16	15	4	0.0011	0.11%
17	16	4	0.0011	0.11%
18	17	2	0.0005	0.05%
19	18	1	0.0003	0.03%
20	**19**	1	**0.0003**	**0.03%**
21	**21**	1	**0.0003**	**0.03%**
Total	---	**3682**	1	100

Ce tableau de couples (n_i, Xi) décrit une distribution statistique. Nous avons vu que le caractère quantitatif prenait habituellement le terme de variable statistique. On parlera de distribution statistique de la variable X.

La détermination des fréquences cumulées

Les fréquences relatives offrent une bonne description de la distribution d'une population en regard d'une variable. Une autre lecture peut être apportée par le cumul des informations. Si l'on reprend l'exemple précédent, un raisonnement en fréquences cumulées permettra de répondre à des questions du type : combien d'individus se déplacent moins de trois fois par jour? combien d'individus se déplacent au moins quatre fois ? combien d'individus ont une mobilité de plus de deux ?...

Pour répondre à ces questions, il sera nécessaire de compléter le tableau précédent. La fréquence cumulée F_i est la somme des fréquences de chaque classe correspondant aux valeurs de la variable statistique X inférieures à l'extrémité supérieure (la borne supérieure) de la classe i :

pour la fonction cumulée croissante : $F_i^c = \sum f_l$ (avec l compris entre **1** et **i-1**)

et

pour la fonction cumulée décroissante : $F_i^d = \sum f_l$ **(avec** l **compris entre i et k)**

k correspondant au nombre de modalités.

En reprenant le tableau relatif à la mobilité, on obtient :

i	xi	ni	fi (%)	Fi croissant	Fi décroissant
1	0	423	11.49%	0.00%	100.00%
2	1	34	0.92%	11.49%	88.51%
3	2	1024	27.81%	12.41%	87.59%
4	3	109	2.96%	40.22%	59.78%
5	4	1042	28.30%	43.18%	56.82%
6	5	217	5.89%	71.48%	28.52%
7	6	466	12.66%	77.38%	22.62%
8	7	119	3.23%	90.03%	9.97%
9	8	119	3.23%	93.26%	6.74%
10	9	45	1.22%	96.50%	3.50%
11	10	28	0.76%	97.72%	2.28%
12	11	15	0.41%	98.48%	1.52%
13	12	17	0.46%	98.89%	1.11%
14	13	8	0.22%	99.35%	0.65%
15	14	3	0.08%	99.57%	0.43%
16	15	4	0.11%	99.65%	0.35%
17	16	4	0.11%	99.76%	0.24%
18	17	2	0.05%	99.86%	0.14%
19	18	1	0.03%	99.92%	0.08%
20	19	1	0.03%	99.95%	0.05%
21	21	1	0.03 %	99.97%	0.03%
Total		3682	100 %	100.00%	0.00%

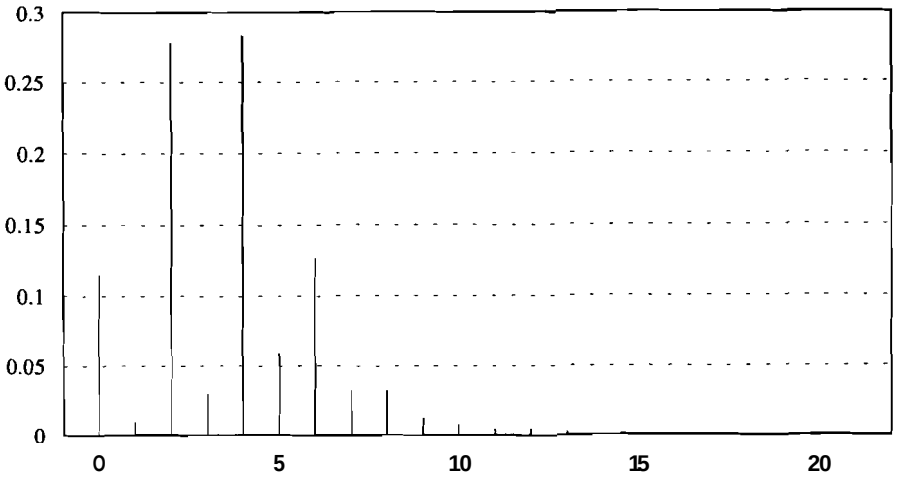
Représentations d'une variable discrète

Le diagramme différentiel

C'est le même principe que pour les variables qualitatives

Répartition selon la mobilité

Diagramme différentiel



Source : enquête ménage Ouagadougou **92**

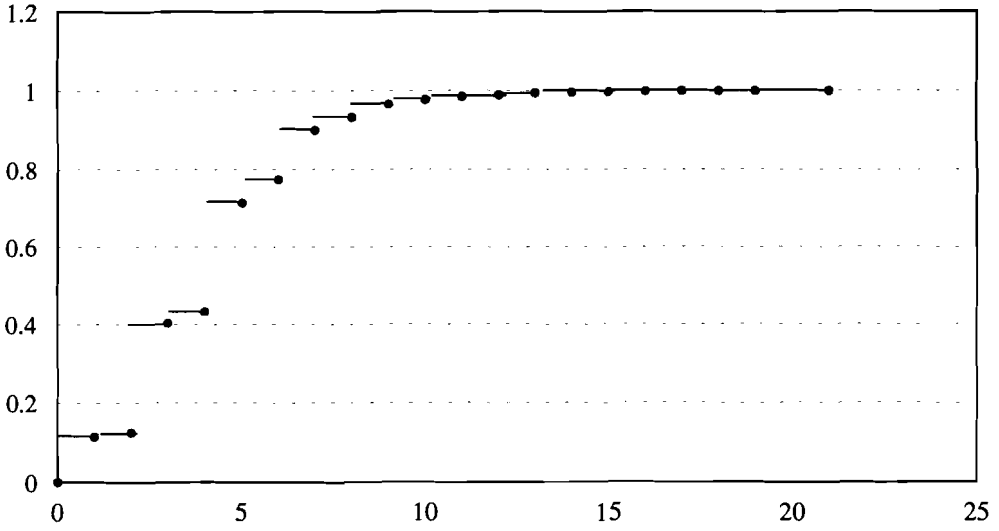
Le diagramme intégral

La représentation de la fonction cumulative conduit à la réalisation d'une courbe en escaliers sur la base de la définition suivante :

$$F(x) = \sum f_l \text{ (1 variant de 1 à i)}, \forall\, x \in\,]x_i, x_{i+1}[$$

Répartition selon la mobilité

Diagramme intégral



Source : enquête ménage Ouagadougou 92

CARACTERE QUANTITATIF CONTINU

La détermination des fréquences relatives

Le principe est le même que pour les variables quantitatives discrètes. Toutefois la constitution de classes intervient de façon beaucoup plus systématique. Le tableau qui suit présente un cas d'application relatif à une variable quantitative continue.

Exemple : répartition de l'échantillon selon l'âge :

x_i	n_i	f_i en %	a_i	f'_i	N_i	F_i
[14-16[	294	8.0	2	12.0	294	8.0
[16-19[	580	15.8	3	15.8	874	23.8
[19-22[	515	14.0	3	14.0	1389	37.8
[22-25[	368	10.0	3	10.0	1757	47.8
[25-31[	590	16.0	6	8.0	2347	63.8
[31-36[	255	6.9	5	4.1	2602	70.7
[36-41[	326	8.8	5	5.3	2928	79.5
[41-51[	407	11.0	10	3.3	3335	90.5
[51-61[	234	6.4	10	1.9	3569	96.9
61 et plus	113	3.1		0.9	3682	100.0
Total	3682	100				

xi sont les différentes valeurs prises par la variable
ni sont les effectifs de chaque classe
fi sont les fréquences relatives
ai sont les amplitudes de classes (différence entre la borne inférieur et supérieure de chaque classe)
f'i sont les fréquences corrigées (nous verrons leur intérêt plus loin)
Ni sont les effectifs cumulés
Fi sont les fréquences cumulées

Représentations d'une variable continue.

La courbe de distribution

La représentation de la distribution d'une variable quantitative continue s'appelle un histogramme. Celle-ci n'est pas sans poser certains problèmes aux étudiants, du moins dans le cas où les amplitudes sont inégales. Prendre en considération l'amplitude des classes, signifie que l'on ne pourra se contenter des hauteurs pour connaître la représentation des individus dans chacune des classes, mais qu'il faudra raisonner sur les surfaces des rectangles. Un telle démarche nécessite de redéfinir des hauteurs de manière à les corriger en fonction de l'amplitude plus ou moins importante. C'est ce nous avons fait dans le tableau précédent dans la colonne des f'i. Pour corriger les hauteurs, on pourra faire appel au simple bon sens, rechercher le plus grand commun diviseur (dans notre cas il s'agit de 3) ou simplement diviser chaque hauteur par l'amplitude puis multiplier, le cas échéant, chaque valeur par un nombre commun permettant de travailler sur des valeurs plus faciles à manier et susceptibles de ne pas alourdir inutilement les graphiques. Il est important de bien comprendre que le seul intérêt des fréquences (ou des effectifs) corrigés est de pouvoir raisonner en terme d'aires. Il ne faudra pas chercher à interpréter numériquement les hauteurs atteintes par les différents

rectangles de l'histogramme et surtout il ne faudra en aucun cas utiliser ces données pour tracer la courbe de répartition.

Source : enquête ménage Ouagadougou 92

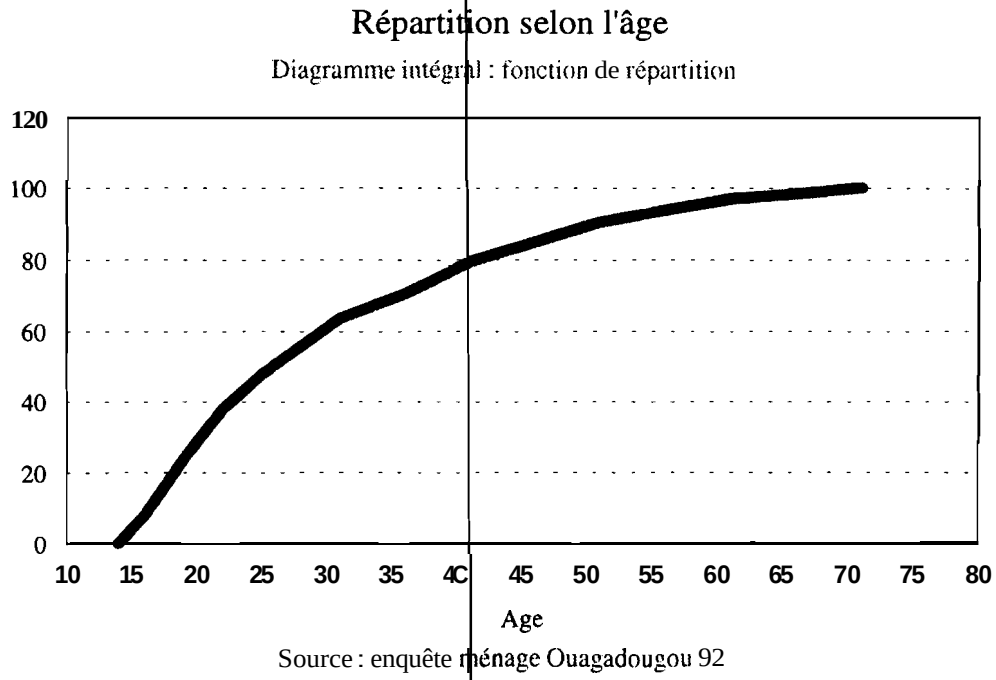
Dernière recommandation importante : les logiciels ont tendance à développer des graphiques qu'ils baptisent histogrammes dès que ceux-ci peuvent être présentés à l'aide de rectangles. Ce terme est abusif dans la mesure ou la représentation offerte par le logiciel correspond davantage aux représentations propres aux variables quantitatives discrètes, ou aux quantitatives continues, lorsque les amplitudes sont égales et que les rectangles apparaissent collés.

Plusieurs problèmes pourront ultérieurement se poser :

- souvent les bornes extrêmes ne sont pas précisées. Elles peuvent être déduites éventuellement si on a une information concernant l'ensemble de la population. Par exemple, si on travaille sur la distribution des salaires dans une entreprise et que l'on dispose de la masse salariale de l'ensemble de l'entreprise, on doit pouvoir définir une borne extrême sans commettre d'erreur trop importante,
- très souvent, on ne dispose pas des données brutes. Lorsque l'on reçoit l'information, la population est déjà répartie en classes. Si l'on souhaite recourir à des instruments statistiques, il est alors nécessaire de poser certaines hypothèses quant à la répartition des individus à l'intérieur d'une classe. L'une des hypothèses les plus courantes, est l'hypothèse d'équirépartition. Cette hypothèse signifie que l'on considère qu'à l'intérieur d'une classe, aux différentes valeurs de la variable correspondent des effectifs identiques. Elle permet également de considérer que la moyenne d'une classe correspond au centre de la classe, c'est à dire à la valeur située à égale distance de la borne inférieure et de la borne supérieure. L'acceptation d'une telle hypothèse n'est pas sans conséquences. Elle pourra dans certains cas surestimer ou sous-estimer certains indicateurs de tendance centrale ou de dispersion.

La fonction de répartition

Le tracé de la fonction de répartition va s'effectuer en prenant pour abscisses les différentes bornes de classes, et comme ordonnées les fréquences relatives ou les effectifs. En aucun cas on ne considérera les fréquences ou effectifs corrigés qui ont servis à dresser l'histogramme.



Source : enquête ménage Ouagadougou 92

2.4. LES CARACTERISTIQUES DE TENDANCE CENTRALES

Un tableau de données ou un graphe apportent une première image d'une distribution. Cependant cette image peut dépendre en partie de la subjectivité de celui qui interprète ces informations. C'est pourquoi pour limiter d'une part cette subjectivité et d'autre part pour en faciliter l'interprétation, on résume les tableaux statistiques au travers de caractéristiques de tendance centrale et de dispersion. Nous allons successivement étudier ces deux ensembles.

Le nombre de données recueillies sur la population étudiée est souvent très élevé. De ce fait ces informations ne peuvent, la plupart du temps, être utilisées comme telles. Pour essayer de caractériser cette population l'on est donc amené à réduire le nombre de paramètres disponibles à ceux nous paraissant les mieux correspondre au problème étudié. On peut par exemple réduire un ensemble de données quotidiennes sur un mois à une donnée moyenne mensuelle.

Plus généralement pour résumer les données recueillies on utilise des indicateurs de tendance centrale tels que le mode, la médiane ou la moyenne.

LE MODE

Le mode d'une distribution correspond à la valeur de la variable pour laquelle **l'effectif** (ou la **fréquence)** est **le plus élevé.**

Cas de variables discrètes

Le mode d'une population peut être déterminé directement à partir d'un tableau de données ou sur la base d'un graphique. Ainsi le mode de la distribution de la mobilité se lit dans le tableau de la page 27 ou sur le graphique différentiel de la page 29 : il est de 4.

Cas de variables continues

Lorsque les classes proposées sont identiques, le mode se détermine aussi simplement que dans le cas d'une variable discrète. On parlera, cette fois, de classe modale. En revanche, lorsque les amplitudes des classes considérées sont inégales, il convient comme pour la détermination de l'histogramme de raisonner sur la base des effectifs ou des fréquences comgées. Si l'on reprend le tableau relatif à l'âge des individus de la page 30, on déterminera la classe modale à partir de la colonne des fréquences comgées. Dans cet exemple la classe modale est la classe [19-22[. Ceci se trouve une fois encore confirmé par l'allure de l'histogramme correspondant situé en page 31.

Intérêts et limites

Intérêts :

- il permet de se faire une première idée de la tendance centrale d'une série,
- connaître sa valeur peut être particulièrement intéressant lorsque la distribution des valeurs est asymétrique,
- il est facile à obtenir,
- il a une signification concrète,
- il permet de repérer des particularités (distribution pluri-modale).

Limites :

- il est très instable,
- il est parfois trompeur (distribution pluri-modale),
- il est mal adapté à l'analyse de distributions statistiques regroupées en classes (variables continues) car sa détermination dépend du découpage des classes,
- il ne se prête pas aux calculs algébriques,
- il est plus sensible que la médiane et que la moyenne arithmétique aux fluctuations d'échantillonnage.

LA MEDIANE

La médiane est la valeur de la variable qui partage les individus statistiques, rangés par ordre de valeurs croissant ou décroissant, en deux sous-populations identiques.

a) Les individus ne sont pas regroupés en classes

Après avoir classé les individus par ordre croissant, par exemple, on déterminera la médiane en se basant sur le rang de l'individu situé au milieu de la population. Cependant l'effectif total de la population peut être pair ou impair.

Si l'effectif de la population est impair, la médiane correspondra à la valeur de la variable pour l'individu situé au rang $(n+1)/2$, n étant l'effectif total.

Si l'effectif total est pair, on s'intéressera aux individus situés aux rangs $n/2$ et $(n+2)/2$. Si ces deux individus présentent une valeur commune par rapport à la variable, on pourra déterminer

une médiane. Dans le cas contraire, il faudra considérer un intervalle médian.

Exemple :

Supposons les notes obtenues lors d'un test, trois cas pourront être distinguées :

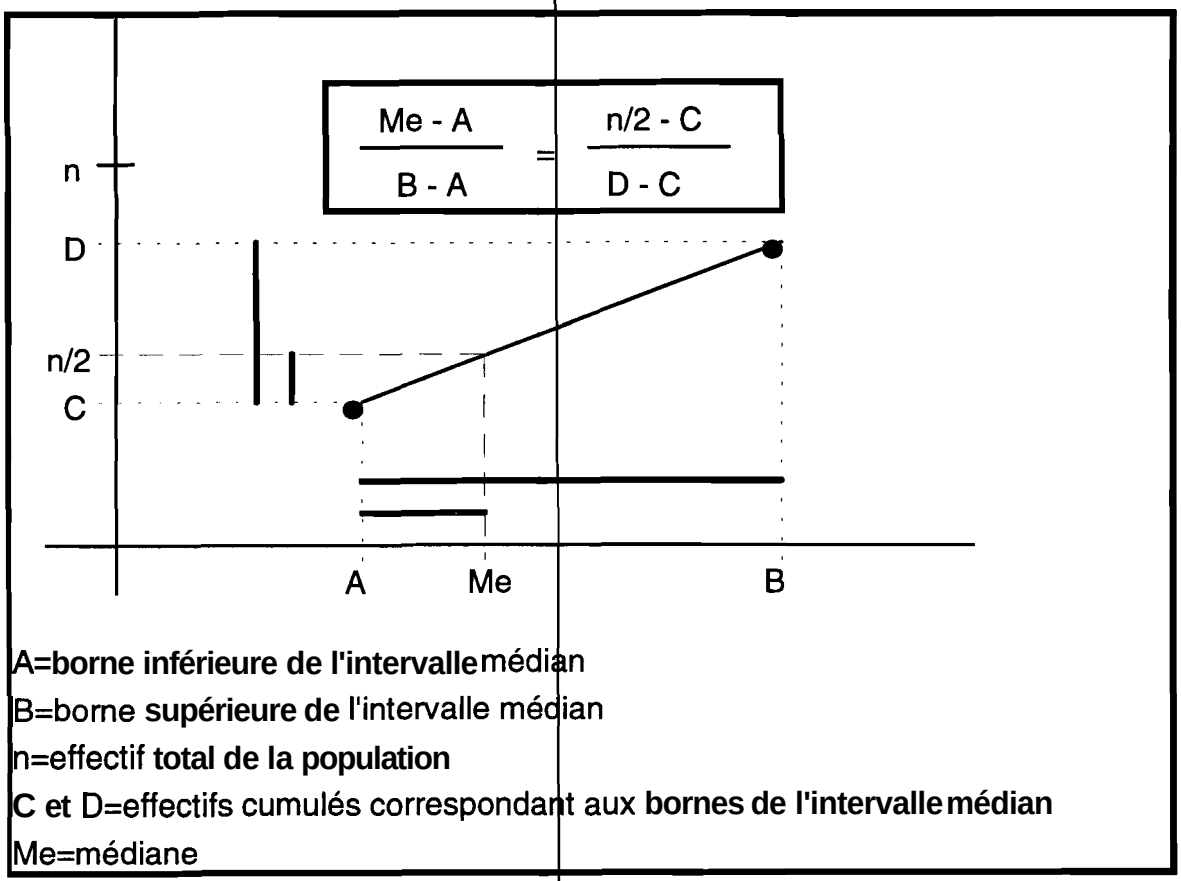
Cas 1 :	8	9	10	11	11	12	12	14	
Cas 2 :	8	9	10	11	11	12	12	13	14
Cas 3 :	9	9	11	11	12	12	12	13	14

Dans le premier cas la médiane est 11, dans le second cas, les individus situés au milieu de la population ont obtenu la même note, la médiane est donc 11. Dans le dernier cas, les individus situés au centre de la série ayant obtenu des notes différentes, on retiendra donc l'intervalle médian [11, 12].

b) Les individus sont regroupés en classes

Pour déterminer la valeur de la médiane on retiendra l'hypothèse d'équirépartition des individus à l'intérieur des classes. On définira, tout d'abord, une classe médiane, puis, le cas échéant, on déterminera par interpolation linéaire une valeur unique pour la médiane. Cette précision peut apparaître, dans certains cas illusoire, du fait de l'hypothèse retenue.

C'est le théorème de Thalès nous permet de déterminer la médiane.



A=borne inférieure de l'intervalle médian
B=borne supérieure de l'intervalle médian
n=effectif total de la population
C et D=effectifs cumulés correspondant aux bornes de l'intervalle médian
Me=médiane

Du graphique précédent on déduit la formule permettant de calculer la médiane :

$$Me = A + [(B - A)*(n/2 - C)] / (D - C)$$

Exemple :

Si l'on reprend l'exemple de l'âge, on constate que la population se compose de 3682 individus statistiques. La classe médiane devra contenir les valeurs prises par les $1841^{ème}$ et $1842^{ème}$ individus. A l'aide de la colonne des effectifs cumulés, on détermine la classe médiane [25-31[.

Il reste à calculer la médiane à l'aide d'une interpolation linéaire :

$$Me = A + [(B - A)*(n/2 - C)] / (D - C)$$

Soit dans notre exemple :

$$Me = 25 + [(31-25)*(3682/2-1757)]/(2347-1757)$$
$$= 25.85$$

Intérêts et limites

Intérêts :

- elle est relativement facile à calculer dans la mesure où il faut simplement classer les observations et cumuler les effectifs,
- elle a une signification claire, précise et concrète,
- elle n'est pas influencée par les valeurs aberrantes.

Intérêts limites :

- son calcul peut poser des problèmes dans le cas où l'hypothèse d'équirépartition pour la classe médiane est douteuse,
- elle se prête moins bien que la moyenne arithmétique aux calculs algébriques et est plus sensible aux fluctuations d'échantillonnage,
- son emploi n'est pas conseillé pour des variables discrètes ayant des "sauts" importants ou des variables continues ne comptant que peu d'observations car sa signification est alors très incertaine.

LES QUANTILES

La logique de la définition des quantiles est la même que celle de la médiane. Au lieu de chercher une valeur qui scinde la série en deux sous-populations égales (50% et 50%), on cherche une valeur qui scinde la population en quatre (quartile), dix (déciles), cent (centiles) ... sous-ensembles égaux.

Les **quartiles** sont les valeurs qui partagent la série en **quatre sous-ensembles égaux.** Ils sont généralement notés : Q1, Q2, Q3. Chaque intervalle qu'ils définissent contient 25% des observations (n/4 de l'effectif n). L'intervalle Q3-Q1 est appelé intervalle inter-quartile et contient 50% des observations.

Les **déciles** sont les valeurs qui partagent la série en **dix sous-ensembles égaux.** Ils sont généralement notés : D1...D9. Chaque intervalle qu'ils définissent contient 10% des observations (n/10 de l'effectif n). L'intervalle D9-D1 est appelé intervalle interdécile et contient 80% des observations.

Les **centiles** sont les valeurs qui partagent la série en **cent sous-ensembles égaux.** Ils sont généralement notés : P1...P99. Chaque intervalle qu'ils définissent contient 1% des observations (n/100 de l'effectif n). L'intervalle P99-P1 est appelé intervalle inter-centile et contient 98% des observations.

LA MOYENNE ARITHMETIQUE

La moyenne arithmétique est l'indicateur de tendance centrale le plus couramment utilisé. Elle correspond au rapport entre la somme des valeurs prises par la variable par les différents individus et le nombre total de ces individus. Nous avons vu que le statisticien pouvait être confronté à plusieurs types de tableaux. De ce fait, deux formules de calcul peuvent être employées suivant l'objectif recherché.

Dans un tableau individus-variables, chaque ligne correspond à un individu et chaque colonne à une variable. Pour calculer la moyenne, il suffit d'additionner les valeurs composant une colonne représentant une variable quantitative, puis de diviser par le nombre total d'individus du tableau. On utilise alors la formule suivante :

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \text{ avec } N = \sum n_i$$

En revanche lorsque les individus sont regroupés en classes, il convient de tenir compte des effectifs, en général différents, de ces classes. On utilise alors la formule de la moyenne arithmétique pondérée, soit :

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{k} n_i x_i$$

ou

$$\bar{x} = \sum_{i=1}^{k} f_i x_i \text{ avec } f_i = \frac{n_i}{N}$$

Par exemple, la moyenne de la mobilité quotidienne du tableau de la page 27 est de :

(0*423+1*34+2*1024+.....+1*19+1*21)=3.71

Celle de l'âge du tableau de la page 30 est de :

(15*294+17.5*580+.....+56*234+66*113)=30.05

Si l'on calcule cette moyenne non pas sur les données regroupées en classe mais sur les données individuelles de départ, nous trouvons 29.66.

Intérêts et limites

Intérêts :

- la moyenne arithmétique tient compte de toutes les observations,
- elle est plus sensible que la médiane aux fluctuations d'échantillonnage,
- elle est de pratique courante,
- elle se prête bien aux calculs algébriques, ce qui explique qu'on la retrouve en statistique inductive ou en analyse de données.

~~Intérêts~~ limites :

- les valeurs exceptionnelles en extrémité de distribution ont une influence sur le résultat,
- pour les variables discrètes, elle ne correspond pas nécessairement à une valeur prise par la variable,
- lorsqu'elle est déterminée à partir de la constitution de classes, il convient de ne pas rechercher de précision illusoire,
- une même moyenne peut provenir de diverses distributions.

2.5. LES CARACTERISTIQUES DE DISPERSION

Les indicateurs de tendance centrale ne rendent compte que d'une partie de la distnbution statistique d'une population. Ainsi, une valeur moyenne de 9 dans un cours n'a pas la même signification si les notes s'échelonnent de 6 à 16 ou si elles varient de 2 à 19... D'où l'intérêt d'étudier la dispersion des données autour de la tendance centrale. Cette dispersion peut être caractérisée par un certain nombre d'indicateurs : l'étendue, l'intervalle inter-quartile, la variance, l'écart-type et le coefficient de variation.

L'ETENDUE

L'étendue est la **différence entre la valeur la plus haute et la valeur la plus basse** de la distribution. Ainsi dans une classe où les notes vont de 6 à 13, l'étendue est de 13-6=7. Lorsque les notes s'échelonnent de 2 à 19, l'étendue est de 17.

L'étendue ne fournit pas une bonne représentation de la dispersion, en effet cet indicateur dépend trop des valeurs extrêmes de la distribution.

L'INTERVALLE INTER-QUARTILE

L'intervalle inter-quartile (Eq) est la **différence entre la valeur du troisième quartile et la valeur du premier quartile.** La méthode de calcul des quartiles a été exposée dans le paragraphe sur les quantiles, dans le chapitre relatif aux indicateurs de tendance centrale.

Pour des séries prenant des valeurs comparables, plus la valeur de l'intervalle inter-quartile est faible, plus la distnbution est regroupée autour de la moyenne. Inversement, un intervalle inter-quartile élevé est le signe d'une grande dispersion.

Les caractéristiques de dispersion de cet indicateur sont assez imparfaites. Elles ne se prêtent que très mal aux calculs algébriques.

Sa détermination est d'autant plus imprécise que l'on se rapproche de valeurs extrêmes : il ne dépend en fait que du rang des observations, et non de leur valeur ou de leur écart relatif. Pour prendre en compte à la fois les valeurs et les écarts relatifs entre observations, il faut utiliser l'écart-type.

LA VARIANCE ET L'ECART-TYPE

L'écart-type est l'indicateur le plus fréquemment utilisé. L'écart-type est la racine carrée de la variance qui se calcule à partir des carrés des écarts à la moyenne que l'on somme et dont on calcule la moyenne.

Lorsaue les individus ne sont pas regroupés en classes

La variance est donnée par la formule suivante :

$$\text{Var} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$ où N est l'effectif total et $\bar{x}$ la moyenne de la série

La variance étant une somme de carrés, les grandeurs qu'elle prend ne sont pas très faciles à interpréter. Pour avoir un indicateur exprimé dans les mêmes unités que les observations, on définit l'écart-type comme étant la racine carrée de la variance :

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Lorsaue les individus sont regroupés en classes

La variance de la série des x_i sera donnée par la formule :

$$\text{Var} = \frac{1}{N}\sum_{i=1}^{N} n_i (x_i - \bar{x})^2$$

ouencore: $\text{Var} = \sum_{i=1}^{N} f_i (x_i - \bar{x})^2$ avec $f_i = \frac{n_i}{N}$

L'écart-type σ, racine carrée de la variance : $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} n_i (x_i - \bar{x})^2}$

ou encore : $\sigma = \sqrt{\sum_{/\,i\,1} f_i (x_i - \bar{x})^2}$ avec $f, = \frac{n.}{N}$

LE COEFFICIENT DE VARIATION

Les indicateurs tels que la moyenne arithmétique ou l'écart-type sont des grandeurs de même espèce que la variable étudiée. Par exemple si les valeurs sont exprimées en francs, il en sera de même pour la moyenne et l'écart-type. Comparer deux séries, dont les observations ne sont pas exprimées dans la même unité par rapport à leur dispersion, peut poser problème. Le coefficient de variation qui est un nombre sans dimension permet de supprimer ces inconvénients. Il est égal à :

$$Cv = \frac{\sigma}{\bar{X}}$$

Là encore, une simple division entre les cellules contenant la moyenne et l'écart-type permettra de calculer le coefficient de variation.

3. L'ANALYSE DES DISTRIBUTIONS A DEUX VARIABLES

Jusqu'à présent, notre travail s'est limité à l'analyse unidimensionnelle des populations qui nous intéressaient. Pour cela nous avons eu recours à des outils nous permettant de décrire les distributions de ces populations, de les représenter graphiquement ou de résumer l'information dont nous disposions à l'aide d'indicateur de tendance centrale ou de dispersion. A chaque fois la population choisie a été observée par rapport à un critère unique. Une telle approche est en général privilégiée au cours de la première phase du traitement d'une enquête. C'est souvent à la suite de cette première analyse que le statisticien va être tenté d'expliquer certains phénomènes qui ont semblé se dessiner. Il va alors s'efforcer au cours d'une seconde étape d'établir ou de vérifier les liens éventuels qui pourraient exister entre certaines variables. Les trois paragraphes qui suivent vont être l'occasion de passer en revue les trois méthodes de base de l'analyse de la distribution à deux variables :

- le tri croisé (croisement de deux variables qualitatives),
- l'analyse de la variance (croisement d'une variable quantitative à expliquer avec une variable qualitative explicative),
- l'ajustement linéaire (croisement de deux variables quantitatives).

3.1. LES DISTRIBUTIONS STATISTIQUES QUALITATIVES BIVARIEES

La présentation des données sous forme de tableaux à deux dimensions (on dit encore ***tableaux*** à ***double entrée*** ou ***tableaux croisés)*** devient nécessaire dès lors que l'on cherche à décrire une population non plus selon un caractère, mais selon deux caractères qualitatifs pris simultanément.

Exemple : la distribution des individus selon le sexe et le niveau d'étude

Sexe \ Niv. Etud.	Ecole coranique	Primaire	Secondaire	Supérieur	Sans étude	Non rép.	Total
Masculin	106	436	861	238	333	10	1984
Féminin	20	347	644	82	603	2	1698
Total	126	783	1505	320	936	12	3682

On peut décomposer cette distribution statistique à deux caractères en plusieurs distributions statistiques simples :

- les distributions marginales, données par les totaux en ligne et en colonne, permettent de connaître la distribution des personnes selon le sexe (total en ligne), et la distnbution des personnes selon le niveau d'étude (total en colonne),
- la distribution des fréquences des personnes selon le sexe et le niveau d'études est obtenue en divisant les effectifs de chaque case du tableau par l'effectif total,

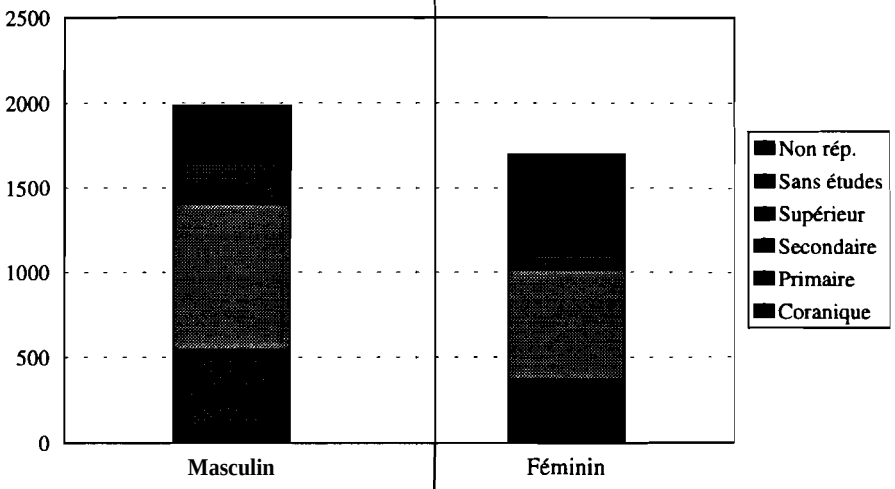
Sexe \ Niv. Etud.	Ecole coranique	Primaire	Secondaire	Supérieur	Sans étude	Non rép.	Total
Masculin	**2.88%**	**11.84%**	**23.38%**	**6.46%**	**9.04%**	**0.27%**	**53.88%**
Féminin	**0.54%**	**9.42%**	**17.49%**	**2.23%**	**16.38%**	**0.05%**	**46.12%**
Total	**3.42%**	**21.27%**	**40.87%**	**8.69%**	**25.42%**	**0.33%**	**100.00%**

- les distributions conditionnelles en ligne et en colonne sont obtenues en divisant respectivement les effectifs de chaque case du tableau 1 par les totaux en ligne et en colonne. Dans notre exemple, la distribution conditionnelle en ligne (ou profil en ligne) représente, pour chaque sexe, la répartition selon le statut matnmonial (tableau 3). La distribution conditionnelle en colonne (ou profil en colonne) représente, pour chaque statut matnmonial, la distribution selon le sexe (tableau 4).

Sexe \ Niv. Etud.	Ecole coranique	Primaire	Secondaire	Supérieur	Sans étude	Non rép.	Total
Masculin	**5.34%**	**21.98%**	**43.40%**	**12.00%**	**16.78%**	**0.50%**	**100.00%**
Féminin	**1.18%**	**20.44%**	**37.93%**	**4.83%**	**35.51%**	**0.12%**	**100.00%**
Total	**3.42%**	**21.27%**	**40.87%**	**8.69%**	**25.42%**	**0.33%**	**100.00%**

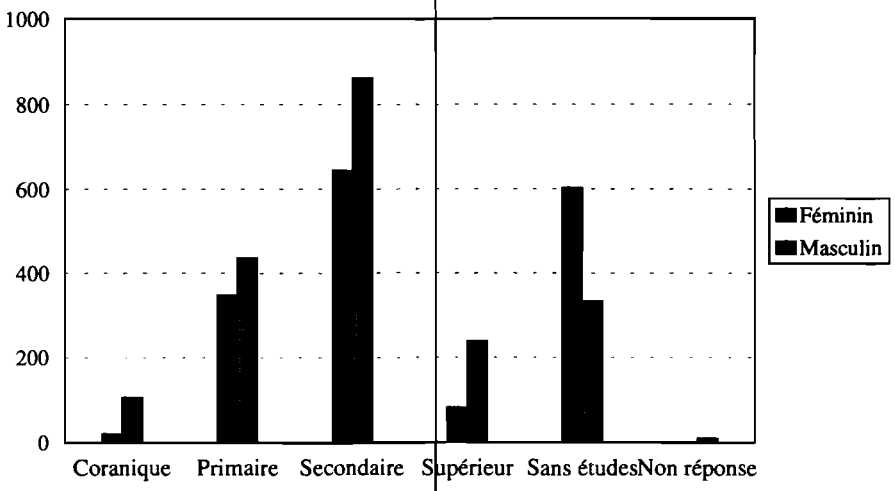
Sexe \ Niv. Etud.	Ecole coranique	Primaire	Secondaire	Supérieur	Sans étude	Non rép.	Total
Masculin	**84.13%**	**55.68%**	**57.21%**	**74.38%**	**35.58%**	**83.33%**	**53.88%**
Féminin	**15.87%**	**44.32%**	**42.79%**	**25.63%**	**64.42%**	**16.67%**	**46.12%**
Total	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**

Répartition selon le sexe et le niveau d'études



Source : enquête ménage Ouagadougou 92

Répartition selon le sexe et le niveau d'études



Source : enquête ménage Ouagadougou 92

3.2. L'ANALYSE DE LA VARIANCE

Nous allons, pour traiter de l'analyse de la variance nous appuyer sur l'exemple d'une étude de la variance de la mobilité en fonction du sexe. L'objectif est de savoir si un découpage de la population en deux catégories , les hommes et les femmes explique la variabilité de la variable **quantitative mobilité.**

Notre premier travail va consister à calculer la mobilité urbaine moyenne des populations masculines et féminines. La moyenne de la mobilité urbaine est de 3,12 pour la population féminine. La moyenne de la mobilité urbaine est de 4,21 pour la population masculine.

Le second est de calculer la variance de la mobilité urbaine. Elle est de 4,80 pour la population féminine et de 6,97 pour la population masculine. Pour la population totale, elle est de 6.25.

LA DECOMPOSITION DE LA VARIANCE

La moyenne est un opérateur linéaire. Elle bénéficie à ce titre d'un certain nombre de propriétés dont l'une des plus importantes s'énonce de la manière suivante : "La moyenne d'une population est une moyenne pondérée des moyennes des sous-populations qui la composent." Cela signifie dans notre exemple que la mobilité moyenne des hommes est égal à la moyenne pondérée des moyennes des hommes et des femmes.

La variance contrairement à la moyenne n'est pas un opérateur linéaire. La variance d'une population n'est pas égale à la moyenne pondérée des variances des différentes sous-populations qui la composent. La variance se compose en fait de deux parties complémentaires que l'on appelle variance inter-classe (variance inter-population - variance expliquée - variance des moyennes) et variance intra-classe (variance intra-population - variance résiduelle - moyenne des variances). Avant de poursuivre, nous allons tenté d'expliquer ces deux composantes.

Lorsque l'on essaie d'expliquer la dispersion des valeurs prises par une variable quantitative comme le salaire par l'appartenance à une modalité d'une variable qualitative comme le sexe, on effectue une décomposition de la variance. Celle-ci va consister à isoler les différentes sous-populations caractérisées par une modalité particulière et à les comparer entre elles.

Il faudra ensuite s'interroger pour savoir si la diversité de la mobilité observée peut s'expliquer par le fait que l'on soit un homme ou une femme ou si, une fois les deux sous-populations isolées, on constate que dans chacun des groupes demeure une forte variabilité de la mobilité. il va donc falloir décomposer la variance observée sur la variable qu'on s'efforce d'expliquer (la mobilité), en deux parties la variance inter-classe et la variance intra-classe, puis mesurer celle qui est la plus importante.

La variance inter-classe

La variance inter-classe mesure la dispersion entre les moyennes de plusieurs sous-populations. Pour mesurer cet écart on va comparer chaque moyenne à la moyenne générale. On va ainsi calculer la différence entre la moyenne d'une sous-population et la moyenne de la population. En ce qui concerne l'exemple qui nous intéresse, on va comparer la mobilité moyenne des hommes avec celle de la population totale, puis la mobilité moyenne des femmes avec celle de la population totale.

Dans la mesure où certains écarts peuvent apparaître positifs, tandis que d'autres seront négatifs, on court une nouvelle fois un risque d'élaborer un indicateur qui dans certains cas fournira des valeurs proches de zéro. Pour éviter ce piège, on élèvera la différence calculée au carré. il restera ensuite à pondérer les écarts mesurés par les effectifs concernés. Il est aisé de constater que cette formule présente de nombreuses similitudes avec la formule de la variance. Il s'agit bien ici du calcul d'une variance des moyennes.

Le calcul de la décomposition de la variance et du pourcentage de variance expliquée dans notre exemple donne :

	Hommes	Femmes	Ensemble
Effectifs	1984	1698	3682
Moyenne	4.21	3.12	3.71
Variance totale	6.97	4.80	6.25
Variance inter-classse	1984/3682*(4.21-3.71)² +	1698/3682*(3.12-3.71)² =	0.29
Variance intra-classe	(6.97*1984)/3682 +	(4.80*1698)/3682 =	5.96

Le pourcentage de variance expliquée est de 0.29/6.25=4.64 %. Ce faible pourcentage signifie que la variabilité de la variable mobilité est plus due à une variance à l'intérieure de chacune des classes qu'à une variance entre les classes. Ceci n'exclut pas le fait que la mobilité peut être très différente selon le sexe.

3.3. L'AJUSTEMENT LINEAIRE

Il s'agit d'étudier le lien entre deux variables quantitatives.
Nous pouvons distinguer deux cas :

- soit les variables X et Y sont indépendantes,
- soit il existe une liaison fonctionnelle et la difficulté est alors de trouver laquelle.

La première étape conduit à réfléchir sur la signification des variables.

La seconde étape consiste à visualiser graphiquement les données. En les représentant dans un repère composé de deux axes, nous obtenons un nuage de points dit "nuage statistique". A l'examen du nuage, on peut avoir une idée du lien de dépendance qui existe entre les variables X et Y. Lorsque le nuage suggère une dépendance relative, on peut se demander dans quelle mesure la connaissance d'une des deux variables permet la connaissance de l'autre. Nous n'étudierons ici que la dépendance linéaire.

LE CHOIX DE LA METHODE

La méthode des moindres écarts conduit à la détermination de la droite qui rend minimum la somme des valeurs absolues des différences :

$$\sum_i |y_i - a.x_i - b|$$

Cette méthode a l'intérêt de donner un poids moindre aux couples $(x_i\ y_i)$ aberrants puisqu'elle fait intervenir les coordonnées à une puissance unitaire mais elle conduit sur le plan algébrique à des calculs inextricables.

La variance intra-classe

La variance intra-classe mesure quant à elle la dispersion qui demeure à l'intérieur de chaque sous-population. Une fois que l'on a isolé les hommes et les femmes, elle va permettre d'établir une synthèse sur la dispersion entre la mobilité d'une femme et la mobilité moyenne des femmes et entre la mobilité d'un homme et la mobilité moyenne des hommes. Une fois calculée la variance propre à la sous-population des femmes, puis la variance propre à la sous-population des hommes, il restera à calculer la moyenne de ces variances en pondérant les variances par les effectifs de chaque sous-population.

Confrontation des variances

Si la variance inter-classe est faible tandis que la variance intra-classe est forte, cela signifie que l'on observe de grandes différences à l'intérieur de chaque classe.

Si au contraire on constate une forte variance inter-classe et une faible variance intra-classe, cela signifie que les individus constituant les sous-populations isolées ont un comportement relativement homogène, tandis que les moyennes de chaque classe sont très hétérogènes entre elles. Dans ce cas on peut considérer que la variable segmentante explique bien la dispersion observée sur la variable quantitative.

Le pourcentage de variance expliquée

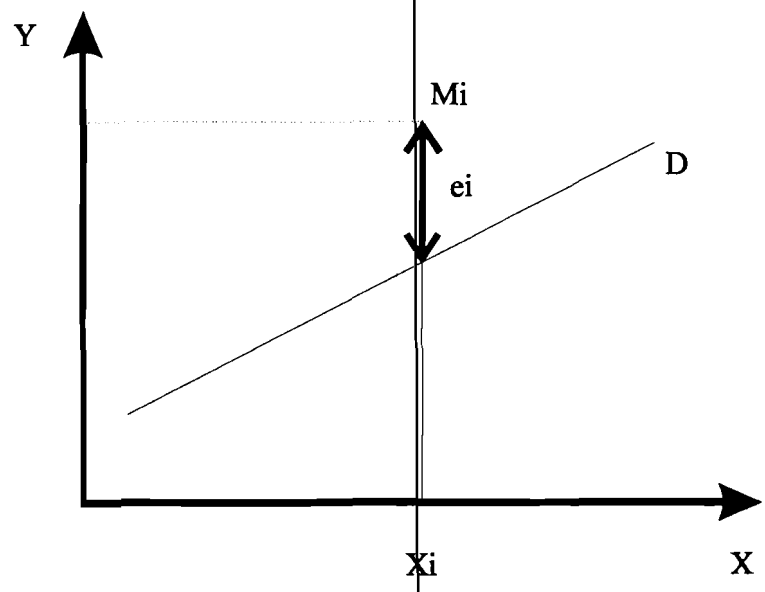
L'indicateur que nous utiliserons pour conclure à la qualité de la segmentation proposée est le pourcentage de variance expliquée. Il consiste à faire le rapport entre la variance inter-classe et la variance totale. Si ce rapport est inférieur à 50 % le pouvoir explicatif de la variable qualitative sélectionnée peut être considéré comme insuffisant.

Dans la mesure où il est extrêmement rare qu'un facteur explique à lui seul toute la variabilité observée, on ne trouvera jamais de pourcentage égal à 100.

Il convient également de se méfier du nombre de classes retenues. Bien souvent, les étudiants sont tentés de conserver un maximum de classes de peur de perdre une partie de l'information. On se retrouve alors avec une quantité considérable de sous-populations. Or, en poussant le raisonnement à l'extrême, ont pourrait très bien envisager d'avoir autant de sous-populations que d'individus dans la population de départ. Que se passerait-il alors ? Il n'y aurait plus aucune variance à l'intérieur de chaque classe. Par conséquent, la variance inter-classe serait égale à la variance totale. Ce n'est pas réellement le but recherché. En fait le travail du statisticien va consister à tenter d'élaborer une typologie en regroupant des individus présentant des comportements homogènes.

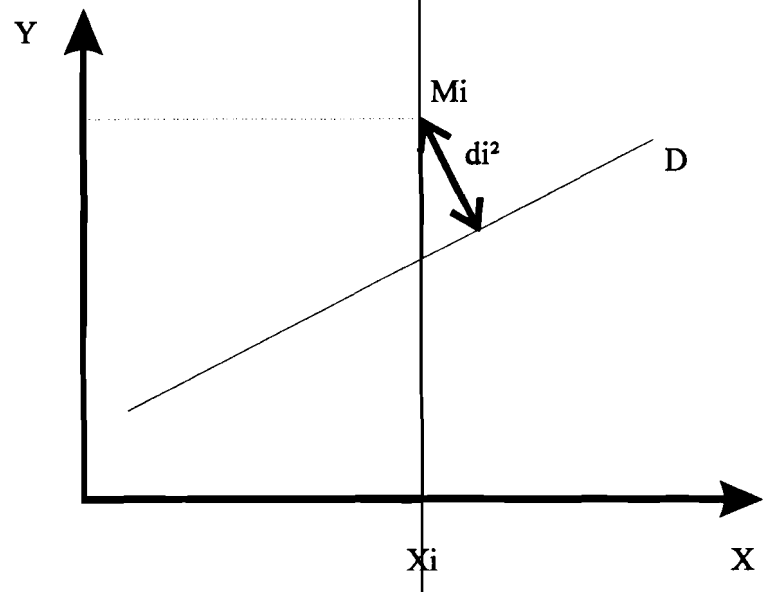
Il n'est pas possible de fournir un nombre limite de classes, car chaque exemple dépend de l'effectif de la population de départ. On s'interdira toutefois de descendre au dessous du seuil de cinq individus par classe. On travaillera ainsi en général sur deux à cinq ou six sous-populations. Cela supposera dans certains cas de procéder à une réagrégation des modalités en acceptant de perdre une certaine richesse d'information au profit d'une plus grande lisibilité.

L'ajustement des moindres écarts



La méthode de l'ajustement orthogonal conduit à déterminer la droite qui rend minimum la somme des écarts comptés orthogonalement à la droite :

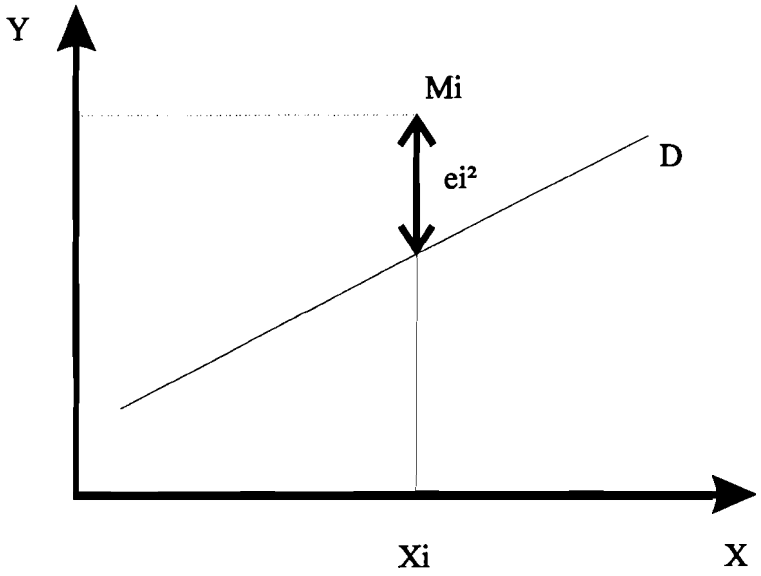
L'ajustement orthogonal



Cette méthode a l'énorme désavantage de fournir une droite sensible aux changements **d'échelle et d'origine.**

C'est pourquoi **la méthode des Moindres Carrés Ordinaires** aboutissant à des droites invariantes par changement d'échelle et d'origine, et de calcul algébrique simple, est la plus utilisée :

L'ajustement des moindres carrés ordinaires



LA DEFINITION DES PARAMETRES

Il s'agit de minimiser les carrés des écarts comptés parallèlement à l'axe des ordonnées, car il est supposé qu'une des deux variables, à savoir l'abscisse X est "prédéterminée" :

Soit la droite D recherchée d'équation $y = a\,x + b$.

Il faut rendre minimum $\sum d_i^2 = \sum (y_i - ax_i - b)^2$

C'est un trinôme du second degré par rapport à b qui est minimum quand la dérivée est nulle, soit :

$$2nb - 2\sum y_i + 2a\sum x_i = 0$$

En divisant par 2n, on tombe sur

$\bar{y} = a\bar{x} + b$ *où* $\bar{x}$ est la moyenne des x et $\bar{y}$ celle des y

Ce qui montre d'une part que $\sum d_i = 0$ et d'autre part que la droite des moindres carrés passe par le point moyen du nuage de points $(\bar{x}, \bar{y})$

En considérant le trinôme précédent par rapport à a. La dérivée s'annule pour :

$$a\sum x_i^2 + b\sum x_i - \sum x_i y_i = 0$$

Ce qui donne si l'on remplace b par $\bar{y} - a\bar{x}$:

$$a = \frac{\sum x_i y_i - n\overline{xy}}{\sum x_i^2 - n\left(\overline{x}\right)^2}$$

Divisons chacun des membres de la fraction par n le nombre d'observations

$$Y = \frac{\sum x_i y_i / n - \overline{xy}}{x_i^2 / n - \overline{x}^2}$$

Vous aurez sûrement reconnu dans la partie inférieure de la fraction, la variance de X. La partie supérieure de la fraction est appelée la covariance.

Nous avons donc a=COV(X,Y)/VAR(X)

La covariance indique le degré de dépendance de deux variables. Si X et Y sont indépendantes alors la covariance est nulle MAIS LA RECIPROQUE N'EST PAS VRAI. Une covariance positive indiquera une relation positive entre les deux variables, c'est à dire que si l'une s'élève, l'autre tendra à le faire aussi. Par contre une covariance négative indiquera une relation négative entre les deux variables, ainsi si l'une est élevée, l'autre tendra à être faible. Le problème de la covariance est qu'elle dépend des unités dans lesquelles les variables X et Y sont exprimées, c'est pourquoi on a cherché à l'améliorer par la définition du coefficient de corrélation :

$$R = \frac{COV(X, Y)}{\sqrt{VAR(X)}\sqrt{VAR(Y)}}$$

Le coefficient de corrélation permet de neutraliser tout changement d'échelle, il varie entre -1 et 1 et mesure l'intensité de la liaison entre X et Y.

R^2 est appelé coefficient de détermination et qualifie la qualité de l'ajustement.

Nous avons donc déterminé la droite d'ajustement avec x comme variable prédéterminée. Cette droite est appelée droite de régression de y par rapport à x. Si Y est considérée comme variable prédéterminée, nous obtenons de la même façon une droite de régression de x par rapport à y, d'équation :

x=a'y+b' où symétriquement

$b' = \overline{x} - a'\overline{y}$ et a'=COV(X,Y)/VAR(Y)

Remarquons qu'en effectuant le produit des coefficients directeurs des deux droites nous obtenons :

$aa'=(COV(X,Y))^2/(VAR(X).VAR(Y))$

soit le carré du coefficient de corrélation R. On a donc $aa'=R^2$.

Attention, l'absence de corrélation linéaire (r=0) n'implique en aucune façon l'absence d'une corrélation non linéaire. Par ailleurs, il faut bien distinguer la liaison corrélative qui concerne des structures de nombres et la liaison réelle qui correspond à une relation de cause à effet. Par exemple, on trouverait une forte corrélation entre l'évaporation d'un liquide et la fermentation d'une matière organique alors que ces deux phénomènes sont totalement indépendants (la chaleur est le phénomène étranger qui a provoqué des variations concomitantes). D'autre part, les couples (x,y) utilisés pour le calcul de la corrélation, ne sont, en général, qu'un échantillon de la population. D'autres échantillons (de même effectif) auraient donné d'autres valeurs de R.

A titre d'exemple, nous donnons les résultats de régressions entre la mobilité et l'âge ou les revenus mensuels chez les habitants de Ouagadougou en 1992. Nous cherchons à expliquer la mobilité par l'âge ou par les revenus. Les équations obtenues sont les suivantes :

Mobilité=-0.027 Age + 4.51 (R^2=0.022)

Mobilité=0.000008 Revenu+3.57 (R^2=0.027)

Le niveau d'explication fourni par la variable explicative est très faible dans les deux cas. Il est surtout surprenant dans le deuxième, car l'on pourrait supposer que la mobilité est fortement liée au revenu.

4. LES ANALYSES FACTORIELLES

Les analyses factorielles, comme d'ailleurs les méthodes de classification abordées ensuite, ne doivent leur vogue actuelle qu'au développement récent de la puissance de calcul des ordinateurs. "Pratiquement", ce sont donc des méthodes récentes, même si leurs fondements mathématiques sont eux connus depuis bien longtemps (calcul matriciel élémentaire).

Les analyses factorielles visent à réduire de manière raisonnée la quantité d'information contenue dans un important d'ensemble de données (de quelques dizaines à plusieurs millions) afin de rendre cette information intelligible. Elles produisent des représentations graphiques simples de la structure initialement invisible des données, tout en cherchant à minimiser les erreurs de perspective dues à l'abandon de certaines facettes des phénomènes.

4.1 LE PRINCIPE GENERAL

Un même principe fonde les diverses méthodes d'analyse factorielle. On a observé N paramètres sur un groupe de K individus et constitué le tableau individus-variables résultant. Ces K individus peuvent alors être considérés comme des points d'un espace à N dimensions. L'inertie de ce nuage par rapport à son centre de gravité, au sens physique (une masse multipliée par le carré d'une distance), représente en fait l'hétérogénéité des points, leur variabilité. Chercher à réduire la masse de données revient ainsi à chercher à réduire la dimension de l'espace dans lequel on observe les individus. Or, normalement (c'est-à-dire hors situation d'indépendance parfaite), le nuage présente des directions privilégiées, sur lesquelles il est plus allongé et donc sur lesquelles l'inertie résultant de la projection des points initiaux est plus élevée. En d'autres termes, le nuage est plus proche d'un ballon de rugby que d'un ballon de football. Des calculs matriciels (il s'agit en fait d'une diagonalisation) montrent qu'il est possible de repérer aisément dans l'espace ces directions privilégiées, appelées axes factoriels. A chacun de ces axes est associé un paramètre quantitatif, la valeur propre, égal à la part d'inertie du nuage de départ reprise par cette direction. On voit ainsi qu'un axe correspondant à une valeur propre faible n'apportera que peu d'information sur le phénomène et pourra donc être délaissé.

Si l'on ne souhaite alors conserver qu'une information par individus, on retiendra la valeur de sa projection sur l'axe correspondant à la valeur propre la plus élevée. Si l'on accepte d'en conserver deux, on ajoutera à cette première direction celle qui correspond à la seconde valeur propre la plus élevée et ainsi de suite. On obtient donc les meilleurs résumés à une, deux, ... dimensions.

Il ne reste plus alors "qu'"à interpréter la signification de ces axes. Des aides à l'interprétation ont été conçues pour faciliter ce travail :

- contribution d'un point au positionnement d'un axe (l'inertie expliquée par l'axe est la somme des inerties des projections sur l'axe des différents points du nuage). Un point à l'inertie élevée sur un axe aura "attiré" cet axe ;
- contribution d'un axe à l'inertie d'un point (on parle souvent de cosinus carré). Elle mesure l'erreur commise lors de la projection du point sur l'axe, erreur décroissant lorsque le cosinus carré se rapproche de 1.

Les différentes méthodes d'analyse factorielle se différencient alors par le type de données qu'elles peuvent supporter.

4.2 L'ANALYSE FACTORIELLE DES CORRESPONDANCES (AFC)

Elle a été conçue originellement pour des tableaux de contingence, mais peut être utilisée avec n'importe quel tableau dans lequel sommes en ligne et en colonne ont un sens. On est alors amené à parler d'analyse factorielle des correspondances multiples (AFCM). L'AFC présente l'originalité de traiter de la même manière les lignes et les colonnes du tableau de départ.

Deux modalités d'une variable sont proches si elles possèdent des distributions similaires sur les modalités de l'autre variable, en d'autres termes si les distributions conditionnelles correspondantes sont proches. L'analyse factorielle des correspondances met donc en évidence des proximités de structure et ne tient pas compte des écarts en valeur. On peut superposer les représentations graphiques des points des deux nuages, mais en gardant alors bien à l'esprit qu'il ne s'agit que d'une commodité de lecture (les deux nuages n'appartiennent en effet pas à des espaces de même dimension). Un point X_i d'un nuage sera alors "proche" d'un point Y_j de l'autre nuage si la modalité j est sur-représentée dans la sous-population i, il en sera éloigné si elle est sous-représentée.

On retrouve donc bien la règle d'interprétation que nous avions énoncée pour les tableaux de contingence : leur étude se fait en comparant les distributions conditionnelles. L'AFC simplifie cette étude en montrant les proximités entre modalités et les raisons de ces proximités.

4.3 L'ANALYSE EN COMPOSANTES PRINCIPALES (ACP)

Si on observe sur une population d'individus les valeurs de N variables hétérogènes (par exemple le prix, la quantité, la distance, ...), les sommes en colonne restent admissibles tandis que les sommes en ligne n'ont plus de sens. L'utilisation de l'AFC multiple reste possible si l'on transforme ces variables quantitatives en caractères qualitatifs. L'analyse en composantes principales permet toutefois de traiter directement ce type de tableau. Un centrage-réduction des variables est le plus souvent indispensable pour éliminer les problèmes d'échelle et de variabilité.

Deux variables sont proches si elles sont corrélées positivement, diamétralement opposées si elles sont corrélées négativement. L'absence de corrélation conduit à des points situés sur des axes perpendiculaires. Deux individus sont proches s'ils présentent pratiquement les mêmes réponses sur

toutes les questions. Contrairement à l'AFC, l'ACP ne révèle donc pas des proximités de structure mais de valeur entre individus.

5. LES METHODES DE CLASSIFICATION

A partir d'une population d'individus caractérisés par un ensemble de variables, les méthodes de classification fournissent des partitions de la population en P classes, les plus homogènes possibles. Contrairement aux méthodes factorielles, on ne cherche pas à ici à réduire le nombre de variables mais le nombre d'individus.

On distingue généralement des méthodes hiérarchiques et des méthodes non hiérarchiques. Les méthodes hiérarchiques produisent une suite de partitions de la population, soit en désagrégeant peu à peu la population (stratégie descendante), soit en l'agrégeant peu à peu (stratégie ascendante). Les méthodes non hiérarchiques fournissent directement une partition de la population en un nombre de classes déterminé à l'avance. Les méthodes descendantes, moins onéreuses en termes de calcul, sont actuellement plus délaissées que les deux autres méthodes.

5.1 LA SEGMENTATION

La segmentation est une méthode descendante hiérarchique. On cherche à expliquer le comportement d'une variable quantitative ou qualitative par un ensemble de variables qualitatives. A la première étape, la partition comprend une seule classe, la population entière. On recherche alors la variable la plus explicative (au sens d'un critère fixé a priori) et l'on teste ensuite toutes les coupures de la population en deux classes permises par des regroupements de modalités de cette variable. La meilleure coupure éclate ainsi la population en deux classes, définies chacune par un ensemble de modalités de la variable explicative. Puis, pour chacun des deux groupes, on reproduit le processus. Pour des variables à expliquer quantitatives, le critère de choix peut être la maximisation de la variance inter-classes (stratégie cohérente avec la technique de l'analyse de la variance). Pour des variables à expliquer qualitatives, on peut retenir la partition fournissant le CHI 2 le plus élevé.

5.2 LA CLASSIFICATION ASCENDANTE HIERARCHIQUE

La classification ascendante hiérarchique (CAH) est une méthode ... ascendante hiérarchique. A la première étape, on dispose d'autant de classes que d'individus dans la population. Soit N ce nombre. On recherche alors les deux individus les plus proches, que l'on agrége. A la seconde étape, on se retrouve donc avec N-1 classes (N-2 classes ne sont d'ailleurs constituées que d'un seul individu). On recherche alors les deux classes les plus proches pour les regrouper. Le processus est répété jusqu'à l'obtention d'une seule classe intégrant l'ensemble de la population. La suite de partitions ainsi créée constitue un arbre, chaque regroupement de deux classes étant dénommé noeud. La proximité entre individus et la proximité entre classes peuvent être définies de diverses manières. Un critère usuel revient à chercher à minimiser à chaque étape la perte d'inertie totale du nuage résultant de l'agrégation de deux classes. Sur le plan pratique, différents algorithmes ont été mis au point afin d'accélérer la vitesse de la méthode (voisins réciproques, ...), mais le résultat final peut dépendre de la stratégie de calcul.

La partition en K classes obtenue par un tel processus est la meilleure partition en K classes, une fois définie la partition en K+1 classes qui est son aînée ; rien ne garantit donc qu'elle soit la meilleure partition en K classes. Enfin, dans le cas où la stratégie d'agrégation retenue est celle de la minimisation de la perte d'inertie, on appelle niveau d'un noeud la perte d'inertie occasionnée par ce regroupement. La somme des niveaux pour l'ensemble des noeuds est alors égale

à l'inertie du nuage. Le choix d'une partition, c'est-à-dire le choix d'un nombre de classes, s'effectue en observant l'allure de la courbe des niveaux cumulés, les décrochages de cette courbe étant des endroits préférentiels de coupure de l'arbre.

5.3 LES NUEES DYNAMIQUES

Les nuées dynamiques sont une des méthodes non hiérarchiques. On se fixe a priori un nombre K de classes et l'on définit une partition initiale en K classes qui fournit K centres de gravité initiaux. On cherche alors à améliorer cette partition en recherchant pour chaque individu le changement de classe faisant décroître le plus possible l'inertie intra-classe. On calcule ensuite de nouveaux centres de gravité et l'on réaffecte alors les individus aux classes définies par ces nouveaux centres de gravité. Le processus est répété jusqu'à satisfaction d'un critère d'arrêt (nombre de mouvements d'individus inférieur à un certain seuil, décroissance de l'inertie intra-classe inférieure à un certain seuil, ...). La partition finale dépend le plus souvent de la partition initiale. On est donc amené à rechercher des "noyaux durs" en identifiant les individus toujours associés dans les partitions obtenues à partir de différentes initialisations du processus, mais avec comme conséquence le risque d'accroître le nombre de classes.

Quelque soit la méthode de classification retenue, on a toujours intérêt, une fois la typologie produite, à caractériser les classes obtenues non seulement à partir des variables de départ mais aussi à partir de variables "illustratives", par exemple de positionnement socio-économique des individus. Plus encore peut-être qu'avec les autres méthodes de la statistique descriptive, les résultats fournis par les méthodes de classification doivent n'être considérés que comme des outils dans une analyse socio-économique.

En conclusion générale sur les méthodes de la statistique descriptive, nous dirons que ces différentes techniques, à une, deux ou de multiples dimensions, sont beaucoup plus complémentaires que concurrentes. Les outils de la statistique descriptive à une dimension servent,dans une première étape, à se familiariser avec les différentes variables. Des liaisons binaires peuvent ensuite être testées par des tableaux de contingence, des analyses de variance, ... Ultérieurement, enfin, on cherchera à réduire le nombre d'individus et de variables en enchaînant analyses factorielles et classifications. Mais il ne s'agit que d'outils et la tâche de l'analyse socio-économique revient toujours à l'analyste, pas à l'ordinateur !

Table de matières