Chapitre 1

La démarche statistique appliquée au management

Minicas

Comment estimer les dégâts d'une catastrophe naturelle (tempêtes, ouragans, etc.) ? Aux États-Unis, la FEMA (*Federal Emergency Management Agency*) a pour mission de prévenir en amont les populations concernées des risques d'une catastrophe naturelle imminente, et de les conseiller sur les actions à mener (brochures avec la liste des principaux conseils, précautions à prendre, etc.). Sa mission est aussi d'aider à la reprise de la vie quotidienne et économique immédiatement après la catastrophe. Dans ce cadre, il est nécessaire que la FEMA puisse évaluer le plus rapidement possible les conséquences de la catastrophe. Elle utilise différents indicateurs, et plus particulièrement des indices mesurant la force des vents (par exemple, l'échelle de Saffir-Simpson), l'hypothèse étant que plus les vents sont violents, plus la zone est dévastée. Un autre indice plus inattendu est utilisé pour estimer le degré de destruction d'une zone : l'indice Waffle House.

Waffle House est une enseigne de 1 600 restaurants spécialisés dans la gaufre. En 2005, après l'ouragan Katrina, sept des restaurants sur la zone sinistrée avaient été détruits, des centaines fermés. En revanche, les restaurants qui avaient pu rouvrir rapidement avaient reçu de très nombreux clients. L'enseigne étant située dans le sud-est des États-Unis, zone où les cyclones sont fréquents et violents, ses dirigeants ont donc décidé d'inscrire dans leur stratégie la gestion de crise liée au passage d'un ouragan et, plus particulièrement, ils ont mis en place tout un dispositif pour que les restaurants puissent ouvrir le plus vite possible : entrepôts situés dans des zones à l'écart de celles qui sont les plus menacées, mode d'emploi de réouverture après le désastre, achats de générateurs portables, d'un centre de commandement mobile, distribution de manuels d'urgence, etc. Ainsi, Waffle House est l'une des entreprises américaines les mieux équipées pour réagir aux catastrophes. Elle est aussi devenue un indice de mesure des conséquences des tempêtes. En effet, au lendemain du passage d'un ouragan, la FEMA estime rapidement l'étendue des dégâts en comptant le nombre de restaurants Waffle House ouverts : lorsque les restaurants sont ouverts et proposent une carte complète, l'indice est au vert, signifiant que les dommages sont limités et que l'électricité fonctionne. Si l'approvisionnement en produits est limité, et l'électricité fournie par un générateur, l'indicateur est au orange. Il est rouge si le restaurant est fermé, signe de graves dégâts ou de conditions météorologiques dangereuses.

Source : *Wall Street Journal*, septembre 2011.

Questions :

1. Quels sont les avantages et les inconvénients de la nouvelle mesure, à savoir l'indice Waffle House ?
2. Comparer les démarches appliquées mises en œuvre pour évaluer les conséquences de la catastrophe dans l'approche météorologique et l'approche par l'indice Waffle House.

Objectifs :

- Comprendre la démarche statistique.
- Décrire les modes de recueil des données.
- Différencier les méthodes d'échantillonnage.
- Identifier les différentes méthodes statistiques.
- Évaluer la qualité d'une démarche statistique.

Chaque jour, un manager est conduit à prendre des décisions pour gérer son entreprise. Leur pertinence dépend directement de la qualité de l'information recueillie en amont de la décision, de sa compréhension, de son analyse et, enfin, de la capacité du manager à transformer l'information en action. Face à une inflation d'informations chiffrées, les managers doivent pouvoir disposer d'outils et de méthodes performants d'aide à la décision : la statistique s'inscrit dans cette perspective.

La *statistique* est un ensemble de méthodes scientifiques dont l'objectif est d'analyser, structurer et modéliser des informations numériques.

Avant de se lancer dans des calculs bien souvent réalisés par des logiciels spécialisés, il convient de réfléchir à la démarche à adopter pour répondre à la problématique managériale soulevée. Ce chapitre s'organise autour des principales étapes de cette démarche statistique. Dans la première section, nous exposons l'étape préliminaire de recueil des informations. Dans la deuxième, nous expliquons comment construire les variables, objets sur lesquels porte une étude statistique. Ensuite, nous présentons les différentes étapes de l'analyse statistique : la troisième section est consacrée à la statistique descriptive dont les méthodes seront développées dans la première partie de cet ouvrage (« décrire ») ; la quatrième section expose les enjeux de la statistique inférentielle qui vise à généraliser à la totalité de la population les résultats obtenus sur un échantillon (méthodes développées dans la deuxième partie de cet ouvrage – « généraliser »). La cinquième section présente la troisième partie de l'ouvrage (« modéliser ») qui porte sur les méthodes explicatives et les méthodes d'extrapolation. Enfin, dans la sixième section, nous insistons sur la nécessité de la mise en place d'une démarche assurant la qualité à tous les stades de l'étude.

1. Recueillir l'information

L'information est la matière première de la statistique. Il est donc nécessaire de la collecter soigneusement. Cette étape est souvent longue et fastidieuse. Elle constitue pourtant le point de départ de l'étude : des bases fragiles aboutissent à un édifice sans robustesse. Dans cette section, nous rappelons l'importance de bien définir la population objet de l'étude (sous-section 1.1), puis indiquons comment construire un échantillon lorsqu'il n'est pas possible d'interroger l'ensemble de la population (sous-section 1.2). Enfin, nous décrivons les quatre principales méthodes de recueil de l'information et les critères de choix (sous-section 1.3).

1.1. Définir précisément la population objet de l'étude

Lors de la construction du problème, le manager doit avant toute chose préciser clairement sur quelle population porte son questionnement et quels individus composent cette population. L'étude porte par exemple sur l'ensemble des entreprises du BTP de moins de 50 salariés et un individu est l'une de ces entreprises.

La *population* est constituée de l'ensemble des individus objets de l'étude.

Un *individu* est une unité de la population.

Un même sujet d'étude peut amener à définir plusieurs populations. Par exemple, Experis France, spécialisée en conseil en recrutement, a lancé en 2011 un baromètre de satisfaction des cabinets de recrutement (http://www.experis-france.fr). Cette étude, menée par TNS-Sofres, est originale car deux études en parallèle ont été réalisées : l'une concerne les directeurs et responsables des ressources humaines, l'autre les cadres âgés de 35 à 50 ans ayant déjà eu affaire à un cabinet de recrutement. L'idée est de comparer les attentes et différences respectives quant à la qualité des prestations des cabinets de recrutement, au suivi de leurs missions et à intégration des nouveaux collaborateurs.

Il faut veiller à ne pas définir une population trop large. Cette situation arrive souvent lorsque le manager raisonne en termes de contraintes budgétaires et souhaite interroger un maximum de personnes avec le même outil de collecte d'information. Cette tentation risque d'avoir des effets très négatifs sur la qualité des résultats. La règle est simple : pour une population spécifique, une étude construite *ad hoc*. En élargissant trop la population, l'outil de mesure risque de ne plus convenir. Par exemple, dans une étude sur la pratique des jeux vidéo, utiliser la même étude pour interroger les jeunes et les parents serait très insatisfaisant : le questionnement n'est pas de même nature. Les jeunes peuvent faire l'objet d'une première étude visant à mieux comprendre leur comportement d'usage (nombre d'heures pratiquées par jour, types de pratique, motivations, etc.). Les parents peuvent être le sujet d'une seconde étude, pour mieux comprendre leurs appréhensions et freins.

Exemple – Étude sur les critères d'achat d'une voiture-taxi

Lors d'une analyse détaillée des ventes en 2010 en France, le directeur du marketing d'un grand constructeur automobile français constate que sa marque est peu présente sur le marché des taxis. Il s'interroge sur les critères de choix de la voiture et les leviers qui pourraient faire évoluer ce choix. Il souhaite également avoir des informations sur le conducteur du taxi (sexe, âge, nombre d'années de pratique, etc.), ainsi que sur son comportement (usage de la voiture, etc.).

Un chargé d'études consulte alors le directeur du marketing pour préciser la cible à interroger. Qu'entend-il par taxi ? Seulement les artisans taxi ? Ou bien considère-t-il également tous les conducteurs de taxi, donc les salariés d'une société de taxi, les locataires, etc. ? Faut-il interroger les taxis dans toute la France ou seulement les taxis parisiens ? Il faut en effet expliciter ces choix dès le début de l'étude.

1.2. Construire un échantillon

Faut-il considérer toute la population ou se limiter à un échantillon extrait de la population ? Dans l'idéal, il conviendrait d'interroger la totalité des individus : c'est l'unique façon de

connaître la réalité des faits. Il n'est cependant pas toujours possible d'obtenir ces informations, notamment en raison de contraintes de temps et d'argent. C'est la raison pour laquelle il est fréquent de réaliser un sondage et de construire un échantillon.

Un *échantillon* est un groupe d'individus extrait de la population.

Construire un échantillon de qualité nécessite des règles précises. Les apprentis sondeurs, grisés par l'usage de nouvelles technologies qui facilitent la remontée des informations, les oublient trop souvent. Par exemple, les médias présentent quotidiennement des résultats de sondages réalisés sur leurs sites, où les internautes peuvent donner leur avis : il faut être bien conscients que ces sondages ne possèdent aucune validité statistique (appelée aussi validité externe), et qu'il ne faut en aucune façon généraliser leurs résultats à l'ensemble de la population. Il en est de même des sites participatifs où sont postées des études faisant appel à des échantillons de convenance : répond qui veut.

Nous présentons ici les deux grands types de méthodes de sondage : les méthodes probabilistes ou aléatoires, puis les méthodes empiriques.

1.2.1. Les méthodes aléatoires de sondage

Dans un sondage aléatoire, chaque individu a une probabilité connue et non nulle d'appartenir à l'échantillon.

Un *échantillon aléatoire* est constitué par un mécanisme aléatoire qui respecte ces différentes probabilités.

Il existe plusieurs méthodes aléatoires. Nous étudions ici les deux principales méthodes : le sondage aléatoire simple et le sondage aléatoire stratifié.

1.2.1.1. Sondage aléatoire simple

Dans le cas d'un *sondage aléatoire simple*, les individus ont tous la même probabilité d'appartenir à la population mère.

Si nous devions réaliser manuellement un sondage aléatoire simple, il faudrait imaginer un tirage de numéros dans une urne comportant tous les noms des individus de la population. Ce type de sondage exige donc de posséder la liste exhaustive de tous les individus qui composent la population, appelée *base de sondage*. Les entreprises des services (banques, assurances, téléphonie, etc.) et les entreprises du e-commerce possèdent la base de données exhaustive de leurs clients et peuvent réaliser facilement des échantillons aléatoires simples en sélectionnant aléatoirement des numéros de clients (tous les logiciels possèdent une fonction aléatoire, par exemple fonction ALEA dans Excel). Dans le cadre des études internes sur les salariés, la base de sondage existe aussi.

On peut réaliser le tirage avec remise ou sans remise : il paraît intuitivement plus satisfaisant de ne pas prendre en compte plusieurs fois le même individu, et donc de construire un échantillon aléatoire sans remise. Dans ce cas, les tirages d'individus ne sont pas indépendants les uns des autres, ce qui complexifie les propriétés et calculs.

Dès que le taux de sondage (la taille de l'échantillon divisée par la taille de la population) est faible (inférieur à 10 %), on peut montrer que les propriétés du tirage sans remise sont proches du tirage avec remise.

Pourquoi le sondage aléatoire simple est peu utilisé ?

Le sondage aléatoire simple est le modèle de référence : la précision de ses résultats sert d'approximation au calcul de la précision des résultats dans d'autres sondages (méthodes des quotas par exemple, voir ci-dessous). Dans la pratique, on utilise toutefois peu ce sondage. Dans de nombreux cas, la base de sondage n'est pas connue : lorsque les achats sont anonymes (par exemple, qui peut donner la liste des consommateurs de lait en France ?) ou lorsque la liste de tous les clients n'est pas disponible. De plus, une population hétérogène sur le phénomène étudié entraîne des résultats obtenus par sondage aléatoire simple peu précis. C'est l'une des raisons pour lesquelles on a développé d'autres méthodes d'échantillonnage, dont le sondage stratifié que nous décrivons maintenant.

1.2.1.2. Sondage aléatoire stratifié

Dans un *sondage aléatoire stratifié*, la population est découpée en plusieurs groupes, appelés strates, puis un tirage aléatoire simple est réalisé dans chacune de ces strates.

Pour construire un échantillon aléatoire stratifié, il convient donc d'identifier au préalable une variable qui permet d'obtenir des strates plus homogènes que la population totale. Il est fortement conseillé de s'appuyer sur des sources documentaires fiables qui mettent en évidence l'influence de cette variable sur la population. Par exemple, les pratiques des directeurs financiers varient beaucoup selon que l'entreprise est petite, moyenne, ou grande : la taille de l'entreprise est une variable de stratification. Lorsqu'une entreprise réalise une étude de satisfaction de clientèle, elle a intérêt à créer des strates selon que les clients sont petits, moyens ou gros.

Les estimations obtenues à partir d'un échantillon stratifié sont plus précises que celles obtenues à partir d'un échantillon aléatoire simple puisque les individus sont plus homogènes à l'intérieur de chaque strate. À condition de bien choisir la variable de stratification !

Une fois les strates construites, on sélectionne les individus aléatoirement au sein de chacune d'entre elles, en respectant le plus souvent les proportions au sein de la population. Par exemple, si la population est composée de 40 % de femmes et de 60 % d'hommes, et que la variable sexe a été utilisée pour la stratification, l'échantillon est composé de 40 % de femmes et de 60 % d'hommes. Il s'agit alors d'un échantillon stratifié proportionnel. Dans certains cas, on peut décider de donner plus de poids à certaines strates et construire un échantillon à probabilités inégales : le poids de la strate dans l'échantillon dépend alors de la taille de celle-ci dans la population et de la dispersion au regard de la variable de stratification.

1.2.2. Les méthodes empiriques d'échantillonnage

Le recours aux méthodes d'échantillonnage non probabilistes ou empiriques est fréquent : ces méthodes ne nécessitent pas de base de sondage et elles sont moins coûteuses. Il n'est pas possible de calculer la précision de ces études comme c'est le cas pour les méthodes probabilistes. Des simulations montrent toutefois qu'elles sont de qualité équivalente lorsque leur protocole est défini de façon précise. Ce sont principalement les sociétés spécialisées dans les études qui les utilisent car elles exigent de l'expertise aussi bien dans la construction de l'échantillon que dans l'administration du questionnaire. L'enquêteur a des consignes précises à respecter.

Dans le cas des *méthodes empiriques*, la sélection des données n'est pas effectuée par sélection aléatoire, mais par un choix raisonné.

Par exemple, dans la méthode des quotas, on construit l'échantillon de façon à représenter un modèle réduit de la population sur deux ou trois de ses caractéristiques essentielles, choisies au regard du thème de l'étude.

Un échantillon construit par la *méthode des quotas* est un échantillon qui respecte la répartition de certaines caractéristiques au sein de la population.

Très classiquement, le sexe, l'âge, la taille du foyer et la profession du chef de ménage sont des critères usuels pour des enquêtes sur des individus physiques. Pour des entreprises, les répartitions selon la taille de l'entreprise, le secteur d'activité et parfois la région sont les critères de quotas. Les critères sont adaptés en fonction du thème de l'étude, mais il est indispensable de connaître leur distribution dans la population. La méthode des quotas implique néanmoins une hypothèse forte, à savoir : si l'échantillon reproduit fidèlement en composition *certaines* des caractéristiques de la population étudiée (sexe, âge, etc.), alors il sera également bon pour *d'autres* caractéristiques non contrôlables, mais qui sont l'objet de l'enquête (par exemple l'intention d'achat, le comportement vis-à-vis d'une entreprise, etc.). Rien ne prouve que cette hypothèse soit vraie.

Du fait d'un abus de langage, certains pensent qu'un échantillon représentatif signifie qu'il est de bonne qualité. Pas nécessairement ! « Représentatif » signifie que l'échantillon « représente » la population dans certaines de ses caractéristiques : ainsi, un échantillon est dit représentatif s'il a été construit par une méthode des quotas.

Exemple – Critères d'achat d'une voiture-taxi : population et méthode de sondage

Il paraît difficile, compte tenu du coût et de la faisabilité, d'interroger l'ensemble des taxis français, même seulement parisiens, et il faudra se contenter de résultats obtenus à partir d'un échantillon.

Le chargé d'études propose de se limiter aux taxis parisiens en considérant toutes les catégories de taxis. Il ne serait pas simple d'obtenir la liste des taxis parisiens, aussi décide-t-il de réaliser un échantillon par la méthode des quotas selon les critères de sexe et du statut (locataire, artisan ou salarié). D'après des informations générales issues d'une étude antérieure, la répartition est de 20 % de femmes et 80 % d'hommes, de 30 % de locataires, de 20 % de salariés et de 50 % d'artisans.

Dans le cas d'un échantillon de taille 500 construit par quota simple, le respect des quotas implique donc d'interroger 100 femmes et 400 hommes, ainsi que 150 locataires, 100 salariés et 250 artisans.

Mais est-on sûr que la répartition femmes/hommes soit la même au sein des trois sous-groupes ? Probablement pas ! Un échantillon par quota croisé respecterait cette répartition à condition, pour construire cet échantillon, de disposer du tri croisé des variables statut et sexe sur la population (voir chapitre 3).

1.3. Les différents modes de collecte de l'information

Il existe quatre grands modes de recueil des données : les méthodes d'entretien, les enquêtes par questionnaire, les observations et les expérimentations.

- Les méthodes d'entretien sont fréquentes en management : quels sont les freins et motivations à faire des achats à partir de son mobile (le m-commerce) ? Sur quels éléments les directions des achats s'appuient-elles pour faire leur choix ? Comment communiquent les dirigeants d'entreprise en cas de crise ? Comment améliorer les conditions de travail ?, etc. Ces méthodes sont menées sur une dizaine d'individus (consommateurs, salariés, responsables d'entreprise, experts, etc.) sélectionnés pour leur diversité d'opinion, de façon à faire émerger les grandes dimensions de la question étudiée. Leur but n'est pas d'interroger, mais d'écouter. Elles sont particulièrement bien adaptées lorsque l'étude demande à explorer, à mieux appréhender, les phénomènes. Ces méthodes sont dites qualitatives, leur vocation n'étant pas de quantifier. Elles ne nécessitent pas ou peu de statistiques : on utilise parfois des méthodes statistiques textuelles dans l'analyse des discours recueillis. Nous ne traitons pas ces méthodes dans le cadre de cet ouvrage.
- Les méthodes d'enquête par questionnaire recueillent des opinions sur un thème donné. Elles permettent de quantifier et de savoir combien de personnes déclarent tel(le) opinion/intention/comportement. On les pratique dans une perspective descriptive ou explicative du phénomène étudié. On peut administrer le questionnaire par téléphone, voie postale, Internet, ou le réaliser en face à face. Les inconvénients de la méthode proviennent essentiellement de la nature déclarative des réponses fournies : de biais propres à la personne interrogée (information fournie incorrecte, volontairement ou involontairement, etc.), ou de biais liés à l'instrument de mesure (question mal posée, mode de collecte inapproprié) ou à l'enquêteur (voir section 6 de ce chapitre). Cette méthode est très utilisée car elle est rapide à mettre en place, et son coût est moyen.
- Par opposition aux méthodes d'enquête par questionnaire, où les individus répondent de façon déclarative, les méthodes d'observation recueillent directement les informations sur le terrain, ce sont des faits. Nous ne traitons pas ici des méthodes d'observation qualitatives (ethnographie, observation participante, etc.), mais souhaitons plutôt parler des méthodes d'observation quantitatives. Il est parfois nécessaire de recourir à un enquêteur pour recueillir les observations : par exemple, des relevés de prix des carburants pour un échantillon de stations-service défini selon les méthodes d'échantillonnage présentées ci-dessus ; le nombre de clients ayant visité un rayon donné, le nombre de produits pris en main, le temps passé dans le rayon (chronométré), etc. Mais on enregistre de plus en plus en souvent l'information automatiquement : par exemple, le temps passé par

un internaute sur un site, le nombre de pages par visite, etc. En finance, on observe le cours d'une action à l'ouverture sur une période donnée, on comptabilise le nombre de retraits d'argent dans le parc de distributeurs automatiques d'une banque, etc. Il faut définir très précisément les mesures observées : par exemple, il faut spécifier si le nombre de visiteurs par heure d'un musée inclut les enfants ou pas, préciser à quelle heure on réalise les relevés, etc. Ces études sont coûteuses, à moins de pouvoir facilement enregistrer automatiquement les données. On les utilise pour décrire des faits. Croisées avec d'autres informations (par exemple, l'âge ou le sexe de l'internaute, la géolocalisation du distributeur automatique, la concurrence, la proximité d'une bouche de métro, etc.), elles permettent d'expliquer le phénomène étudié (par exemple, nombre de sites visités, nombre de retraits d'argent, etc.).

- Les expérimentations permettent de manipuler certains facteurs et de tester leur impact. Par exemple, est-il plus efficace d'adresser un courriel de prospection dont l'objet est personnalisé et comporte le prénom du destinataire, ou pas ? L'expérimentation consiste à créer deux échantillons : l'un reçoit le courriel avec l'objet personnalisé, l'autre le courriel dont l'objet est standard. Un nouveau procédé de production permet-il de raccourcir les délais de fabrication ? On réalise un test comparant l'ancien et le nouveau procédé. Ces méthodes ont pour avantage de mesurer l'efficacité directement. Si on les exécute dans des conditions proches de la réalisation finale, elles permettent d'obtenir des résultats très précis. Toutefois, réaliser une expérimentation demande souvent du temps et la méthode est coûteuse. La complexité du protocole d'expérimentation (appelé aussi plan d'expérimentation) augmente avec le nombre de facteurs testés. On n'utilise ces méthodes que dans une perspective explicative d'un phénomène.

Le choix de la méthode de collecte dépend de nombreux critères, dont les plus importants sont les suivants :

- la nature du problème managérial : explorer, décrire ou expliquer ; le tableau 1.1 récapitule quel mode est le plus adapté à quelle problématique ;
- les contraintes budgétaires : par exemple le face à face à domicile est très onéreux, mais parfois obligatoire lorsque le questionnaire est long ou complexe (montrer des prototypes par exemple) ;
- les contraintes de temps : certaines méthodes demandent des délais de réalisation plus longs que d'autres ; par exemple, une expérimentation où l'on mesure l'impact d'une publicité télévisée et de ses caractéristiques sur une zone test demande au moins six semaines ;
- le nombre d'informations à collecter et leur complexité : l'appréciation de la nouvelle texture d'une crème impose un questionnaire en face à face s'il faut tester le produit ; des images peuvent s'afficher facilement sur des enquêtes en ligne ;
- la dispersion géographique souhaitée : un recueil exigeant un face à face n'est pas réalisable si l'étude nécessite une forte variété géographique.

Tableau 1.1 : Méthode de collecte de l'information et nature du problème managérial

	Explorer	Décrire	Expliquer
Entretien	√		
Enquête par sondage		√	√
Observation		√	√
Expérimentation			√

2. Construire les variables statistiques

Une fois le mode de collecte fixé, la deuxième étape consiste à construire et caractériser les objets statistiques étudiés.

Dans la sous-section 2.1., nous définissons la notion de variable statistique. Nous différencions ensuite les deux grands types de variables, quantitative et qualitative (sous-section 2.2.), avant de décrire soigneusement ces deux catégories (sous-sections 2.3. et 2.4.) et de présenter le tableau individus/variables, point de départ de l'analyse statistique (sous-section 2.5). Nous expliquons enfin pourquoi l'identification du type de variable est essentielle (sous-section 2.6).

2.1. Notion de variable statistique

Un ensemble de critères pertinents au regard de l'étude décrivent chaque individu : nous définissons ainsi les *variables statistiques*.

Le tableau 1.2 illustre quelques exemples de populations, individus et variables.

Tableau 1.2 : Exemples de populations, individus et variables statistiques

Population	Individu	Variables statistiques
Ensemble des abonnés d'un opérateur de téléphonie	Un abonné	Âge, sexe, services possédés, montant de la facture mensuelle, etc.
Entreprises de BTP	Une entreprise	Nombre de salariés, chiffre d'affaires, délais de paiement, etc.
Usagers d'un restaurant d'entreprise	Un usager	Emploi, âge, sexe, fréquentation, dépense, appréciation, etc.
Ensemble des pièces produites par une chaîne de production	Une pièce	Diamètre, poids, robustesse, etc.

2.2. Variables qualitatives et quantitatives

On distingue les variables statistiques qualitatives des variables statistiques quantitatives. Le tableau 1.3 donne quelques exemples.

Une *variable statistique quantitative* est une variable associée à un caractère mesurable.

Une *variable statistique qualitative* est une variable associée à un caractère qui n'est pas mesurable.

Tableau 1.3 : Exemples de variables qualitatives ou quantitatives

Variable	Type
Catégorie socioprofessionnelle de l'usager	Qualitative
Chiffre d'affaires de l'entreprise	Quantitative
Nombre d'enfants	Quantitative
Année de naissance	Qualitative

Exemple – Critères d'achat d'une voiture-taxi : variable qualitative ou quantitative ?

Le statut (par exemple, artisan, salarié, locataire, etc.) est une variable qualitative. Le prix de la dernière voiture-taxi achetée est une variable quantitative. L'âge mesuré en nombre d'années est une variable quantitative. En revanche, si l'âge avait été mesuré en tranche, la variable aurait alors pu être considérée comme qualitative.

Les variables mesurant l'importance des critères d'achat (prix de la voiture, marque, etc.) peuvent être qualitatives ou quantitatives selon la façon dont elles sont formulées. Ainsi, si la personne interrogée indique par exemple une note comprise entre 0 et 10, la variable est considérée comme quantitative. Si le répondant exprime son appréciation en choisissant entre différentes possibilités (par exemple, pas du tout important, pas important, important, très important), alors la variable est qualitative.

On peut éventuellement identifier par des nombres les modalités prises par une variable qualitative (par exemple les codes de départements ou les années). Il ne s'agit pas pour autant d'une variable quantitative : effectuer des opérations arithmétiques sur ces nombres ne fait pas sens (on ne peut pas additionner des codes de département !).

2.3. Variables qualitatives nominales ou ordinales

Les réalisations possibles d'une variable qualitative s'appellent des modalités. Par exemple, la variable sexe présente deux modalités : femme/homme (voir tableau 1.4).

Lorsque les modalités d'une variable qualitative peuvent se ranger selon un ordre précis, la variable est dite qualitative ordinale (ou ordonnée). Par exemple dans le cas d'une échelle d'usage, on peut considérer la modalité « jamais » comme inférieure à la modalité « quelques fois » qui, elle-même, est inférieure à la modalité « souvent ». Autrement, la variable est dite qualitative nominale, par exemple le lieu de résidence dont les modalités sont les départements d'habitation.

Une *variable statistique qualitative nominale* est une variable dont les modalités ne peuvent pas être classées selon un ordre préétabli.

Une *variable statistique qualitative ordinale* est une variable dont les modalités peuvent être classées.

Tableau 1.4 : Exemples de variables qualitatives et de modalités

Variable	Modalités possibles	Variable qualitative nominale/ordinale
Catégorie socioprofessionnelle	Ouvrier qualifié, agent de maîtrise, ingénieur, etc.	Nominale
Sexe	Femme, homme	Nominale
Jugement	Très insatisfait, insatisfait, satisfait, très satisfait	Ordinale

2.4. Variables quantitatives discrètes ou continues

Les réalisations possibles d'une variable quantitative correspondent à un ensemble de valeurs.

Une variable quantitative est dite discrète lorsqu'elle prend un nombre limité de valeurs entières. C'est le cas par exemple du nombre d'enfants (voir tableau 1.5). Lorsqu'une variable quantitative prend un nombre de valeurs qu'il n'est pas possible de dénombrer, alors la variable est dite quantitative continue (par exemple le cours d'une action qui varie à tout moment).

Une *variable quantitative discrète* prend un nombre limité de valeurs entières.

Une *variable quantitative continue* est une variable qui peut prendre toutes les valeurs dans un intervalle donné.

Tableau 1.5 : Exemples de variables quantitatives et de valeurs possibles

Variable	Valeurs possibles	Variable quantitative discrète/continue
Nombre d'enfants	0, 1, 2, etc.	Discrète
Note de satisfaction (sur 10)	0, 1, 2, …, 9, 10.	Discrète
Prix payé pour un repas	5,5 €, 6,8 €, etc.	Continue

2.5. Le tableau individus/variables

Une fois les variables construites et les informations collectées, les données sont rassemblées dans un tableau où chaque ligne récapitule l'ensemble des informations concernant un individu M_i, et chaque colonne présente l'ensemble des informations concernant une variable X_j (voir tableau 1.6).

Le *tableau individus/variables* reporte les valeurs ou les modalités prises par les N individus pour les p variables.

Tableau 1.6 : Tableau des individus/variables

Variables / Individus	X_1	…	X_j	…	X_p
M_1	x_{11}	…	x_{1j}	…	x_{1p}
…	…	…	…	…	…
M_i	x_{i1}	…	x_{ij}	…	x_{ip}
…	…	…	…	…	…
M_N	x_{N1}	…	x_{Nj}	…	x_{Np}

Lecture : x_{ij} est la valeur prise par l'$i^{ème}$ individu M_i pour la variable X_j.

2.6. Pourquoi la nature de la variable est-elle aussi importante ?

La nature de la variable est fondamentale. Tout d'abord, elle permet de déduire les méthodes statistiques à appliquer. En effet, certaines méthodes sont adaptées pour le traitement de variables qualitatives, d'autres pas. Dans la suite de cet ouvrage, pour chaque méthode présentée, nous spécifierons systématiquement le type de variables concerné.

Ensuite, d'un point de vue statistique, il est préférable de collecter des variables quantitatives continues sur lesquelles on peut appliquer des méthodes plus nombreuses et plus sophistiquées. Toutefois, dans la pratique, il n'est pas toujours possible de collecter des variables quantitatives continues : on ne demande par exemple jamais le montant annuel des revenus d'un ménage, variable essentielle pour comprendre le comportement des consommateurs, sous la forme d'une variable quantitative continue. Le risque de non-réponse ou de réponse volontairement falsifiée est élevé. Il est ainsi préférable de proposer des tranches de revenus : cela signifie que la variable est transformée en qualitative ordinale.

Outre le fait de contourner en partie le risque de non-réponse, dans le cas de l'exemple précédent, l'intérêt d'une variable qualitative ordinale est de pouvoir affecter une valeur à chaque modalité en respectant l'ordre dans les modalités. Par exemple, pour l'échelle de satisfaction, plutôt que d'inclure une variable qualitative nominale où les deux modalités sont oui/non, il est préférable d'inclure une variable qualitative ordinale d'échelle d'accord. Au lieu de « êtes-vous satisfait, oui/non ? », il est conseillé de proposer une échelle de satisfaction, c'est-à-dire de passer d'une variable qualitative nominale à une variable qualitative ordinale : « Très insatisfait » peut prendre la valeur 1, « insatisfait » la valeur 2, « ni insatisfait ni satisfait » 3, « satisfait » 4 et « très satisfait » 5. Bien sûr, cette affectation est arbitraire, mais elle permet néanmoins de calculer une note moyenne de satisfaction. La variable qualitative ordinale est ainsi rendue quantitative discrète. Aussi est-il nécessaire de considérer la nature de la variable dès sa formulation.

3. Décrire l'information : la statistique descriptive

Les données ont été collectées, elles sont prêtes à être analysées. Pour cela, la statistique a développé de nombreux outils de description des données.

La *statistique descriptive* a pour objet de résumer et de présenter l'information contenue dans des données collectées sur un groupe d'individus.

Les outils de la statistique descriptive varient selon que l'analyse concerne une, deux ou plusieurs variables (analyse univariée, bivariée et multivariée).

3.1. Description variable par variable, la statistique univariée

Considérons les variables une par une et analysons l'information du tableau de données. Il s'agit de décrire comment varie une variable donnée *Xj*, c'est-à-dire d'analyser comment se répartissent ses valeurs ou modalités (x_{1j} ; … ; x_{Nj}).

La *statistique descriptive univariée* fournit les outils statistiques pour organiser, présenter et synthétiser l'information issue de l'analyse d'une variable indépendamment des autres.

La statistique descriptive univariée est un premier niveau d'analyse incontournable. Elle permet de prendre contact avec les données. Les informations sont présentées à partir de tableaux, de diagrammes, et d'indicateurs de tendance et de dispersion qui synthétisent les résultats. Nous décrivons au chapitre 2 les principaux outils de la statistique univariée appliquée au management.

Exemple – Critères d'achat d'une voiture-taxi : utilisation de la statistique univariée

Par exemple, l'analyse descriptive univariée des données collectées sur les taxis permet de connaître l'âge moyen des conducteurs de taxis, le nombre moyen d'années d'expérience, les principaux critères d'achat de leur véhicule, d'ordonner ces critères d'achat du plus important au moins important, de quantifier combien ont bénéficié d'un crédit pour effectuer l'achat, le pourcentage de taxis ayant été conseillés dans leur choix d'une nouvelle voiture par des collègues, etc.

3.2. Lien entre deux variables, la statistique bivariée

Les premières conclusions apportées par la statistique univariée mènent à envisager d'autres questions portant sur l'analyse croisée des informations recueillies.

Exemple – Critères d'achat d'une voiture-taxi : utilisation de la statistique bivariée

Le modèle de la voiture-taxi est-il lié au statut du taxi (indépendant, salarié, etc.) ? Le prix payé à l'achat varie-t-il en fonction de l'âge ?

La *statistique descriptive bivariée* a pour objet d'étudier conjointement deux variables *X* et *Y* sur une même population.

Elle vise à appréhender les relations qui peuvent exister entre les variables. Nous exposons au chapitre 3 les principales méthodes de la statistique bivariée. Elles varient selon la nature des deux variables (qualitatives et/ou quantitatives).

L'analyse bivariée permet de mettre au jour des phénomènes qui peuvent ensuite faire l'objet d'explorations plus avancées ou conduire à envisager un modèle explicatif.

3.3. Analyse simultanée des variables, l'analyse multivariée

Les méthodes d'analyse utilisées au cours des deux premières étapes sont des techniques incontournables de dépouillement des données. Elles ne permettent toutefois pas de traiter des tableaux dans leur globalité.

La *statistique multivariée* vise à étudier plusieurs variables simultanément.

Parmi les méthodes multivariées descriptives, l'analyse des données multidimensionnelles met à la disposition du manager des méthodes d'analyse descriptive globale permettant de faire ressortir l'information principale contenue dans une grande quantité de données.

Le choix de la méthode utilisée varie en fonction du type de variables considérées : analyse en composantes principales (ACP) dans le cas où toutes les variables sont quantitatives, analyse factorielle des correspondances (AFC) dans le cas d'un tableau de contingence croisant deux variables qualitatives, analyse des correspondances multiples (ACM) dans le cas de variables qualitatives.

Les résultats obtenus à l'aide de ces méthodes apparaissent sous forme de graphiques qui peuvent donner une impression trompeuse de simplicité. Il est indispensable que le manager soit en mesure de comprendre et d'interpréter les résultats dans toute leur complexité.

Nous consacrons le chapitre 4 à l'analyse en composantes principales. Sans développer les bases mathématiques complexes de ces méthodes, nous donnons les clés pour interpréter les sorties d'un logiciel d'analyse des données et éviter les erreurs d'interprétation.

Les méthodes de classification font aussi partie des méthodes multivariées descriptives. Leur objectif est de regrouper les individus en un nombre limité de classes homogènes. On utilise fréquemment ces méthodes en marketing (par exemple, les segments de clientèles), en ressources humaines (profils managériaux), etc. Les méthodes de classification font l'objet du chapitre 5.

Exemple – Classification des consommateurs et consommation durable

Le cabinet Ethicity, en collaboration avec Aegis Media, réalise une typologie annuelle des consommateurs français sur leurs comportements en matière de consommation durable. Il suggère également des leviers d'action aux entreprises. La figure 1.1 reproduit les résultats obtenus dans l'étude de mars 2011.

En termes de consommation responsable, Ethicity constate tout d'abord que les consommateurs français se différencient avant tout en fonction de leur pouvoir d'achat et de leurs convictions. Ce sont les deux axes représentés sur la figure 1.1.

L'analyse a permis d'identifier trois grands groupes de consommateurs, répartis en huit sous-groupes. Par exemple, les bio-beaux (14 %). Ils sont davantage centrés sur eux-mêmes, ils recherchent des produits de qualité et des solutions pour leur bien-être. Ils sont acteurs du changement (pour le développement durable) pour des considérations de santé. Ils sont les plus hédonistes et assument une consommation plaisir et saine qui participe à leur santé en général (ils n'ont pas réduit leur consommation et ne sont pas dans le boycott). Les éco-restreints (15,6 %) s'intéressent au développement durable, mais en privilégient les comportements qui leur permettent de faire des économies (moins de gaspillage, produits faits maison, etc.). Ils appartiennent en général aux classes moyennes ou modestes de la population.

Parmi les consommateurs sceptiques, on trouve les minimiseurs (17,9 %) qui estiment en faire déjà suffisamment à travers certains gestes quotidiens, comme le tri des déchets, et ne veulent pas faire davantage d'efforts. Pour eux, la reprise de la croissance économique est un enjeu bien plus important que le développement durable, et ils ne sont donc pas prêts à modifier leur consommation.

À partir de cette typologie, le cabinet a identifié des leviers à activer pour inciter les consommateurs français à aller vers une consommation plus responsable. Par exemple, pour consophages, Ethicity conseille de montrer les astuces et bons plans, d'adopter une modernité du discours *via* le digital. Pour les perméables, il faut privilégier les bénéfices sociaux, aller vers eux, pratiquer une information très simple, notamment *via* les étiquettes.

Les autres résultats se trouvent sur http://www.blog-ethicity.net/

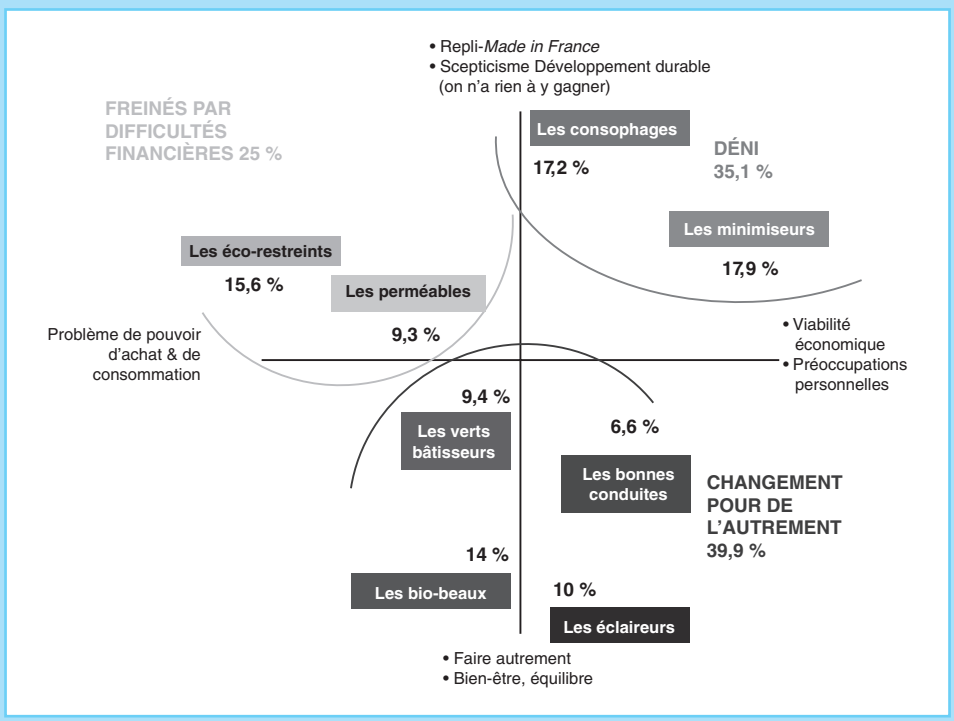


Figure 1.1

Dimensions de la consommation durable et typologie des Français.

Source : Ethicity, 2011.

4. Généraliser les résultats obtenus à partir d'un échantillon à l'ensemble de la population : la statistique inférentielle

4.1. Le principe de la statistique inférentielle

Souvent, le groupe d'individus décrit au cours de l'étape exploratoire ne constitue pas la population complète sur laquelle porte l'étude, mais un échantillon issu de cette population (voir sous-section 1.2 de ce chapitre). La question se pose alors de la généralisation des résultats trouvés sur l'échantillon à la population complète. C'est l'objet de la statistique inférentielle.

La *statistique inférentielle* consiste à décrire la population à partir d'observations faites sur l'échantillon. Les caractéristiques inconnues d'une population sont déduites à partir d'un échantillon issu de cette population.

Afin de distinguer indicateurs sur la population et indicateurs sur l'échantillon, nous identifions systématiquement, dans cet ouvrage, les premiers par des lettres grecques et les seconds par des lettres romaines.

4.2. Estimer un résultat à partir d'un échantillon et évaluer sa précision

D'un côté, nous nous intéressons à un phénomène sur une population, mais les valeurs des paramètres – l'objet même de l'étude, par exemple le stock moyen mesuré en nombre de pièces disponibles par magasin, le taux de satisfaction des clients d'un opérateur de téléphonie mobile, ou bien l'appréciation par les clients de la gestion des réclamations – sont inconnues car impossibles à calculer. De l'autre côté, nous possédons des valeurs connues de ces paramètres sur un ensemble d'observations issues de la population, un échantillon. La statistique inférentielle propose des outils qui permettent d'extrapoler les résultats obtenus sur un échantillon à la population dont est issu cet échantillon (voir figure 1.2).

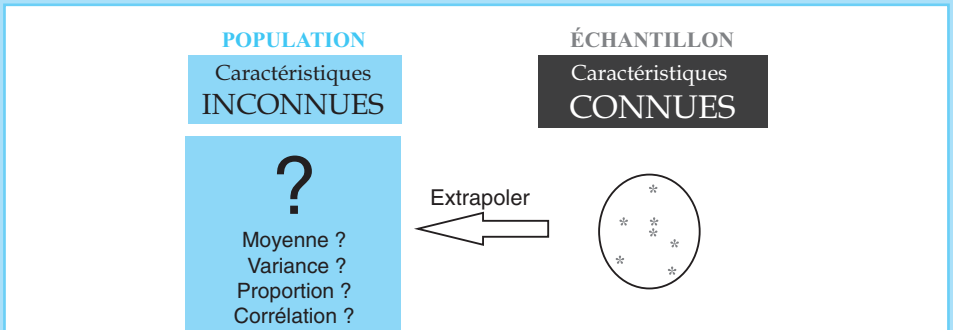


Figure 1.2

Objet de la statistique inférentielle : comment généraliser les résultats issus de l'échantillon vers la population ?

Cela soulève de nombreuses questions auxquelles nous allons répondre dans le chapitre 6, consacré aux méthodes d'estimation : premièrement, comment estimer la vraie valeur sur la population ? Quels indicateurs donnent les meilleures estimations ? Deuxièmement, comment extrapoler vers la population les résultats obtenus sur l'échantillon ? Quelle est la précision des résultats ? Fatalement, on commet une erreur en ne tenant pas compte de la totalité de la population : peut-on quantifier cette erreur, appelée erreur d'échantillonnage ?

4.3. Prendre une décision grâce à la statistique

Au-delà du fait d'attribuer une valeur à un paramètre sur la population, le problème est parfois de répondre à une question concernant la population. Le lien entre dépense annuelle et sexe, observé sur un échantillon de clients d'un site de e-commerce est-il vrai pour l'ensemble des clients de ce site ? Les pièces fabriquées par un fournisseur respectent-elles la norme fixée dans le cahier des charges ? Le taux de satisfaction est-il supérieur chez les clients résidant en Île-de-France ou chez ceux habitant en régions ?

Pour répondre à des questions posées sur une population à partir de résultats sur un échantillon, on met en œuvre une procédure de test d'hypothèse. La question de l'erreur commise en prenant une décision à partir de résultats d'échantillon est, là aussi, incontournable (en fait, deux types d'erreurs sont identifiés).

Le chapitre 7 expose la méthodologie des tests d'hypothèses. Le chapitre 8 décrit six tests d'hypothèse fréquemment utilisés en management : deux tests de comparaison à une norme, trois tests de comparaison de populations et un test d'association.

Exemple – Qualité d'un fichier

> Posséder une base de données comportant de nombreux clients et contacts, c'est bien ! Mais disposer de données de qualité, c'est encore mieux ! Or les données collectées sont rapidement obsolètes et leur mise à jour est coûteuse en temps et en argent. Plus particulièrement, l'adresse postale exige une actualisation régulière. Le SNA (Service National de l'Adresse) estime à 1 euro le PND (Pli Non Distribué) du fait d'une erreur dans l'adresse. Au total, les PND représentent une perte annuelle de 183 millions d'euros pour les entreprises, 7 000 tonnes de papier gaspillé, soit 300 fois la hauteur de la tour Eiffel, ou encore une fois le tour de la Terre (source : www.laposte.fr/sna).
>
> Alors, comment juger de la qualité d'un fichier et de la récence de ses adresses avant d'exécuter une campagne en grand nombre ? Un test sur un échantillon aléatoire extrait de la base de données permet de s'assurer que le taux de retour PND est inférieur à une norme admise dans le métier du marketing direct, à savoir 5 %. Au-delà de ce chiffre, il faut lancer d'urgence une procédure de mise à jour des données (par exemple, en achetant Charade, le fichier des déménagés commercialisé par la Poste) avant que la base de données ne devienne inexploitable.

5. Quantifier les relations entre des variables : la modélisation

La dernière étape de la démarche statistique n'est pas systématique et ne concerne que les études les plus avancées. L'objectif est de *modéliser*, c'est-à-dire de quantifier les relations entre des variables.

Un *modèle statistique* est une simplification de la réalité qui vise à formaliser des relations entre plusieurs variables.

Par exemple, en marketing, se pose souvent la question de l'efficacité sur les ventes des dépenses publicitaires et de l'impact du prix. Existe-t-il un lien entre ces variables ? Si oui, quelle est sa nature et son intensité ? Grâce à la formalisation des relations, le modèle permet de répondre à des questions telles que : si le budget publicitaire augmente de 10 %, quel sera l'impact sur les ventes ? Combien vaut l'élasticité au prix, c'est-à-dire si le prix baisse de 10,%, dans quelle proportion la demande augmente-t-elle ?

Un organisme financier (banque, affactureur, etc.) peut vouloir expliquer le taux de défaillance des entreprises par des indicateurs financiers (par exemple, fonds de roulement, bénéfices, ventes, avoir des actionnaires, etc.) et déterminer quels indicateurs de liquidité ou profitabilité expliquent fortement le taux de défaillance. Existe-t-il des indicateurs qui pourraient être observés plusieurs mois avant la faillite, et donc aider à l'anticiper ?

Il s'agit de trouver une formule mathématique décrivant le plus précisément possible la réalité. Deux familles de modèles existent : les modèles explicatifs (sous-section 5.1) et les modèles d'extrapolation (sous-section 5.2), présentés maintenant.

5.1. Les modèles explicatifs

Les méthodes explicatives visent à déterminer s'il existe une relation entre une variable à expliquer notée Y (la défaillance d'une entreprise, par exemple) et une ou plusieurs variables explicatives notées X_i (taille de l'entreprise, secteur d'activité, indicateurs de solvabilité, liquidité ou profitabilité, etc.), et à quantifier la relation le cas échéant.

Les *méthodes explicatives* s'attachent à déterminer une fonction f qui modélise la relation liant les p variables explicatives $X_1, X_2, \ldots, X_p$ et Y soit $Y = f(X_1, X_2, \ldots, X_p)$.

La relation entre les deux variables est orientée, dans le sens où les variables $X_1, X_2, \ldots, X_p$ « expliquent » la variable Y. Il est important de ne pas confondre association et causalité : les méthodes explicatives n'établissent pas une relation causale, mais l'existence d'une association entre les variables. Seule la théorie est susceptible de définir une relation de causalité. Le bon sens aussi est souvent très utile : entre la température extérieure et les ventes de glaces d'un fabricant, laquelle est la variable explicative ? Nous laissons le soin au lecteur de répondre ! Dans certains cas, la réponse est plus complexe. Ainsi, on peut légitimement penser que le prix explique le niveau des ventes. Maintenant, les entreprises fixent aussi leurs prix en fonction du niveau des ventes et, par exemple, les baissent si les ventes n'ont pas atteint leurs objectifs : les deux variables sont donc liées par une relation de cause à effet. Ce problème dit d'endogénéité est fréquent en management. Il n'est pas simple à résoudre : nous n'abordons pas dans cet ouvrage les méthodes intégrant l'endogénéité.

Il existe de nombreuses méthodes explicatives : régression simple, régression multiple, régression logistique (Logit/Probit), analyse de variance, analyse discriminante, etc. Les méthodes se distinguent en fonction de la nature qualitative ou quantitative des variables explicatives et à expliquer. Nous invitons le lecteur à se reporter au tableau 1.7 : celui-ci croise

variables à expliquer et explicatives selon leur nature, puis récapitule les principales méthodes explicatives associées.

Dans le cadre de cet ouvrage, nous approfondissons deux méthodes explicatives très utilisées en management : la régression simple (chapitre 9) et la régression multiple (chapitre 10).

Tableau 1.7 : Quelle méthode explicative choisir ? Tableau récapitulatif (non exhaustif) des méthodes selon la nature des variables à expliquer et explicatives

		Variables explicatives		
		Nominale(s)	**Quantitative(s)**	**Nominale(s) & quantitative(s)**
Variable à expliquer	**Nominale**	Régression logistique (Logit /Probit)	• Régression logistique • Analyse discriminante	Régression logistique
	Ordinale	• Analyse conjointe • Logit/Probit ordonné	Logit/Probit ordonné	Logit/Probit ordonné
	Quantitative discrète	Régression de Poisson	Régression de Poisson	Régression de Poisson
	Quantitative continue	Analyse de variance (ANOVA)	Régression simple ou multiple	• Analyse de covariance (ANCOVA) • Régression multiple

Exemple – Quelle méthode explicative choisir ?

- Comment expliquer le taux d'attrition (*churn*), c'est-à-dire le pourcentage de clients qui se désabonnent au bout d'un an, par des variables sociodémographiques (sexe, âge, revenu, etc.) ou bien comportementales (gros client ou, au contraire, client occasionnel, catégories de produits achetés, plaintes formulées, etc.) ? Une régression logistique mettra en relation l'état du client (a résilié/n'a pas résilié) avec l'ensemble des variables explicatives nominales et métriques citées ci-dessus.
- Comment déterminer si un client va renouveler son équipement (par exemple son téléphone mobile, son parc d'ordinateurs ou d'imprimantes, son automobile, etc.) ? Une analyse discriminante vise à trouver les variables (ancienneté de l'équipement, âge, revenus, montant des dernières factures, nombre de produits possédés, etc.) et leur pondération de façon à calculer pour chaque client un score d'appétence.
- Comment un constructeur automobile peut-il définir la voiture électrique de demain ? À côté du travail de R&D des ingénieurs, le département marketing doit s'assurer des préférences du consommateur et définir les caractéristiques de l'offre en termes de taille, de consommation, d'autonomie, de couleur, de prix, etc. Une analyse conjointe permet de relier des préférences (produit préféré au produit le moins apprécié) avec les caractéristiques du produit.
- Le taux de service (pourcentage de commandes livrées à temps par une entreprise) varie-t-il entre plusieurs sites de production ? Une analyse de variance est adaptée.

- Une société d'assurances souhaite connaître le profil de ses assurés en fonction du nombre de leurs accidents passés, de leur âge, du sexe, du type de voiture, de la région d'habitation, etc. La variable à expliquer est quantitative discrète (nombre d'accidents) et une régression de Poisson s'applique.
- Impact du marketing (montant du budget publicitaire par type de support [TV, Internet, presse, etc.], niveau de prix, budget accordé aux actions promotionnelles, prix pratiqué par les principaux concurrents) sur le niveau des ventes : toutes les variables sont quantitatives et il faut réaliser une régression multiple.

5.2. Les méthodes d'extrapolation

Prévisions des ventes, prévisions de la trésorerie à trois mois ou à six mois, prévisions des volumes d'achat de matières premières pour la production, prévisions des cours des matières premières, etc. : les besoins en prévisions dans une entreprise sont fréquents. Celles-ci nécessitent la mise en œuvre de méthodes d'autant plus robustes qu'elles sont à l'origine de décisions managériales essentielles : recrutement de nouveaux salariés, investissements pour de nouvelles chaînes de production, définition des besoins en financement, etc. Les méthodes d'extrapolation répondent à ce besoin de prévision.

Elles partent du principe qu'un phénomène observé dans le passé va se reproduire dans un futur proche. On effectue donc la modélisation par rapport au temps. On obtient ensuite la prévision en extrapolant le passé vers le futur (d'où la dénomination de ces méthodes).

Les *méthodes d'extrapolation* consistent à regarder la forme du phénomène observé dans le passé, puis à la projeter dans le futur.

Les méthodes statistiques d'extrapolation le plus souvent utilisées sont la méthode des moyennes mobiles, le lissage exponentiel et la méthode de Box et Jenkins. Nous consacrons le chapitre 11 à la méthode des moyennes mobiles.

Exemple – Prévision des revenus SMS/MMS : illustration du principe des méthodes d'extrapolation

L'ARCEP (Autorité de Régulation des Communications Électroniques et des Postes) est l'autorité administrative indépendante en charge de la régulation des télécommunications et du secteur postal. L'une de ses principales missions consiste à analyser l'évolution des marchés et à « veiller à ce que la concurrence s'exerce effectivement » (source : www.arcep.fr). Pour mener à bien sa mission, l'ARCEP a mis en place plusieurs observatoires et produit les données officielles sur le marché des télécommunications : nombre de lignes fixes, nombre d'abonnements aux services mobiles, volume de téléphonie, revenus issus des services mobiles, etc.

La figure 1.3 présente les recettes trimestrielles provenant des services SMS/MMS (tous opérateurs confondus), en million d'euros, observées entre le 1er trimestre 2001 (noté T1 2001) et le 2e trimestre 2011 (noté T2 2011). Quelles seront les prévisions de recettes pour les trimestres 3 et 4 de 2011 ? La forme de phénomène observée est très proche d'une droite. La méthode d'extrapolation consiste à poursuivre cette droite dans le futur pour obtenir des prévisions. Nous prévoyons ainsi (le modèle est développé sur le site compagnon) une recette de 658,1 millions d'euros pour le 3e trimestre 2011 et de 671,7 millions d'euros pour le 4e trimestre 2011. En avril 2012, l'ARCEP a publié les vraies valeurs pour les 3e et 4e trimestres : elles sont de 671 et 682 millions d'euros, ce

qui signifie que le modèle d'extrapolation utilisé est de très bonne qualité : l'erreur de prévision est inférieure à 2 %.

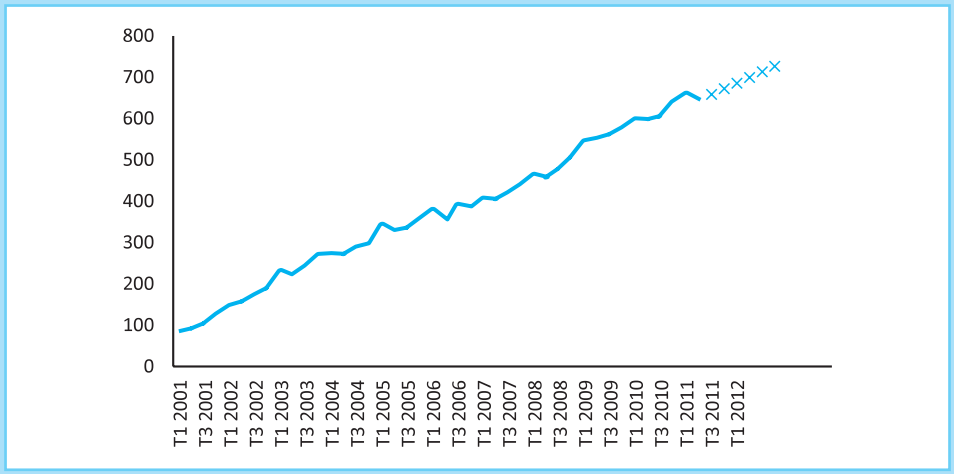


Figure 1.3

Revenus des services SMS/MMS (en million d'euros) entre le 1er trimestre 2001 et le 2e trimestre 2011 et prévisions.

Source : ARCEP.

6. Réaliser une démarche statistique de qualité

Dans cette dernière étape, il s'agit de réfléchir humblement à la qualité de l'étude statistique réalisée. Il faut admettre que, dans de nombreux cas, il est difficile d'appliquer une démarche scientifique rigoureuse en management : les différentes disciplines du management sont complexes, de nombreuses variables interagissent, les actions des concurrents et du marché provoquent des chocs externes inattendus et difficilement maîtrisables, les concepts sont parfois imprécis et, pourtant, on souhaite les mesurer (par exemple, qu'est-ce qu'un salarié fidèle ? Un salarié satisfait ? Un salarié dans l'entreprise depuis longtemps ?, etc.).

Ainsi, des erreurs susceptibles de rendre les résultats caducs et inexploitables peuvent survenir à toutes les étapes de l'étude. Décrivons-les maintenant.

6.1. La qualité dans la collecte de l'information

À quoi bon mettre en œuvre des méthodes statistiques avancées si les données collectées initialement sont de mauvaise qualité et ne permettent pas d'appréhender le phénomène observé ? Les principales sources d'erreur d'une étude statistique au stade de la collecte d'informations sont l'erreur de couverture et l'erreur de non-réponse.

L'*erreur de couverture* provient d'une différence entre la population cible à étudier et la population réellement étudiée. Autrement dit, la population étudiée ne couvre pas la totalité de la population que l'on souhaiterait étudier. Par exemple, dans les études sur la fréquentation des sites sur Internet, un logiciel doit être téléchargé sur l'ensemble des ordinateurs habituellement utilisés par le répondant, qu'ils soient personnels ou professionnels. Si l'étude arrive à bien couvrir les ordinateurs personnels, elle couvre en revanche mal les ordinateurs professionnels car l'installation de logiciels est souvent interdite par les services informatiques des entreprises.

L'*erreur de non-réponse* provient de l'absence partielle ou complète d'informations concernant les individus de l'échantillon. L'erreur de non-réponse est grave et biaise les résultats de l'étude si les non-répondants sont atypiques et ont un avis différent des autres. Mais, comment le savoir ?

Une non-réponse partielle importante est souvent preuve d'un problème dans l'outil de mesure : la question est mal formulée, la rendant difficilement compréhensible, le thème abordé est confidentiel, la question demande à retrouver des éléments qui n'ont pas été archivés, le questionnaire est trop long, etc.

Il faut toujours s'interroger sur les raisons pour lesquelles des individus ne répondent pas une enquête, et travailler soigneusement l'outil de collecte de l'information (par exemple, le questionnaire) pour qu'il soit le plus simple et le plus compréhensible possible, et ne provoque ni rejet total ni rejet partiel.

Que faire avec les non-réponses ?

Si le taux de non-réponse est trop élevé, il est parfois préférable d'annuler la question. Si la non-réponse partielle concerne quelques individus, on peut réaliser les analyses statistiques en mettant de côté ces individus pour la question à laquelle ils n'ont pas répondu. On peut aussi imputer une valeur à la réponse manquante : la règle est arbitraire et l'on peut remplacer la valeur manquante par la moyenne sur l'ensemble de la population interrogée ou bien par la valeur d'un individu ayant le même profil. Des méthodes plus sophistiquées existent. Quoi qu'il en soit, ce n'est toujours qu'un pis-aller.

Exemple – Les enquêtes de satisfaction des salariés et l'anonymat

Pour avoir une vision des attentes et comportements des salariés, les directeurs des ressources humaines conduisent régulièrement des études de satisfaction des salariés de l'entreprise : comment perçoivent-ils leur encadrement ? Quel est leur niveau d'engagement et de motivation ? Quelles sont leurs plus grandes sources de frustrations ? Le risque dans ces études est la non-réponse et des réponses biaisées par peur de représailles si les identités étaient dévoilées. Par conséquent, un atout clé dans les enquêtes de satisfaction des salariés est le respect de l'anonymat. Cela doit se traduire à tous les niveaux de l'enquête : dans le mode de collecte (proscrire le téléphone et préférer le questionnaire autoadministré reçu au domicile du salarié), dans la lettre d'accompagnement, dans les consignes adoptant un ton rassurant et affirmant un engagement de confidentialité de la part de l'entreprise, dans la forme des questions (éviter les questions qui permettraient d'identifier facilement la personne, essayer d'adopter une formulation neutre), etc. Malgré cela, ne soyons pas dupes : des non-réponses et des réponses biaisées demeurent.

6.2. La qualité dans la mesure de l'information

Lors d'études passées, on a trouvé que le nombre d'ordinateurs déclarés achetés était supérieur au niveau des ventes d'ordinateurs (une donnée réelle connue des fabricants et revendeurs).

Un écart entre les réponses enregistrées et les vraies valeurs s'appelle une *erreur de mesure.*

La question posée était pourtant bien anodine : avez-vous acheté un ordinateur au cours des 12 derniers mois ? La formulation n'était pas en cause. Après réflexion, on est arrivé à la conclusion que les répondants avaient eu tendance à acquiescer plus qu'ils n'auraient dû : en effet, certains souhaitaient donner une bonne image d'eux en déclarant posséder un produit impliquant socialement (biais de désirabilité sociale), d'autres ne se souvenaient pas parfaitement de la date d'achat (qui avait eu lieu au-delà des 12 mois). Dans ce cas précis, la vraie valeur du nombre d'ordinateurs achetés étant connue, l'erreur a pu être redressée mais, dans la majorité des cas, la vraie valeur est inconnue (c'est justement l'objet de l'étude !).

Les erreurs de mesure proviennent essentiellement de trois sources :

- du répondant : celui-ci n'est pas capable de répondre (problème de mémoire, réponse demandée trop précise, etc.) ou bien il falsifie sa réponse (biais psychologique) ;
- de l'instrument de mesure : il est trop long, la question est mal formulée, le mode de recueil est inadapté, etc. ;
- des enquêteurs : ils peuvent inconsciemment biaiser dans un sens (par exemple, par une façon non neutre de poser les questions).

Que faire pour réduire les erreurs de mesure ?

Il faudrait systématiquement organiser des prétests : faire relire le questionnaire par d'autres personnes, passer le questionnaire sur un petit échantillon en condition réelle, etc. Ce sont des actions simples qui permettent d'anticiper et de réduire la non-réponse et des réponses biaisées. Malheureusement, pour des raisons de manque de temps, cette étape est souvent négligée. Le suivi et la formation des enquêteurs constituent aussi un moyen de réduire l'erreur de mesure.

6.3. La qualité d'un résultat obtenu sur un échantillon

En généralisant à l'ensemble de la population un résultat obtenu sur un échantillon issu de celle-ci, nous ne pouvons avoir la certitude que le résultat observé sur l'échantillon correspond à la vraie valeur sur la population. Cette erreur due à l'échantillonnage n'est pas due à une mauvaise sélection de l'échantillon, comme certains le pensent.

L'*erreur d'échantillonnage* provient des fluctuations dues au principe même de l'échantillonnage.

On peut, sous certaines conditions, quantifier cette erreur. Le chapitre 6 est consacré à la mesure de l'erreur d'échantillonnage.

6.4. L'écart entre modèle et réalité

Comme nous l'avons vu (section 5), un modèle est une simplification de la réalité. La qualité d'une étude statistique dépend donc aussi de la capacité du modèle à approcher la réalité.

L'*erreur de modélisation*, aussi appelée *erreur d'ajustement*, provient de l'écart entre le modèle et la réalité.

Cette erreur est mesurable : il s'agit de la différence entre les données observées et les données ajustées par le modèle. Dans les chapitres 9, 10 et 11, nous calculons les erreurs d'ajustement dans le cadre des modèles de régression simple, de régression multiple et de décomposition par moyennes mobiles.

6.5. La qualité globale d'une démarche statistique

La qualité globale d'une démarche statistique dépend de toutes ces erreurs qui s'accumulent au fur et à mesure de l'étude. L'erreur d'échantillonnage s'ajoute aux erreurs de couverture, de non-réponse, de mesure et d'analyse pour constituer l'erreur totale.

À l'exception de l'erreur d'échantillonnage et de celle de modélisation, il n'est pas possible de mesurer explicitement la valeur des erreurs. Seule une évaluation de la qualité à travers un examen de la démarche statistique entreprise permet de conclure si les erreurs sont maîtrisées dans leur ensemble. Comme l'illustre la figure 1.4, il ne sert à rien d'agir pour réduire un certain type d'erreur, si les autres restent très élevés : par exemple, augmenter la taille de l'échantillon pour réduire l'erreur d'échantillonnage a un impact faible sur la qualité des résultats si l'erreur de mesure est élevée (par exemple, un outil de mesure défaillant).

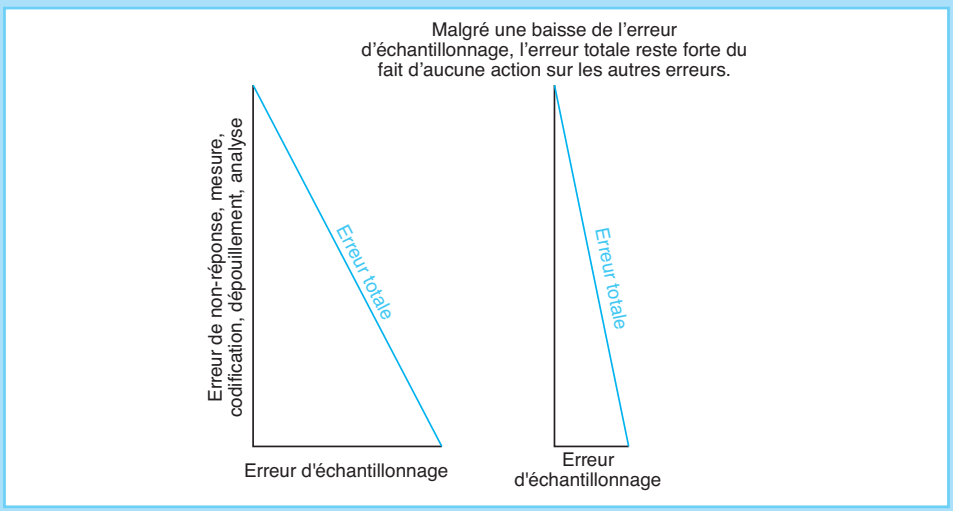


Figure 1.4

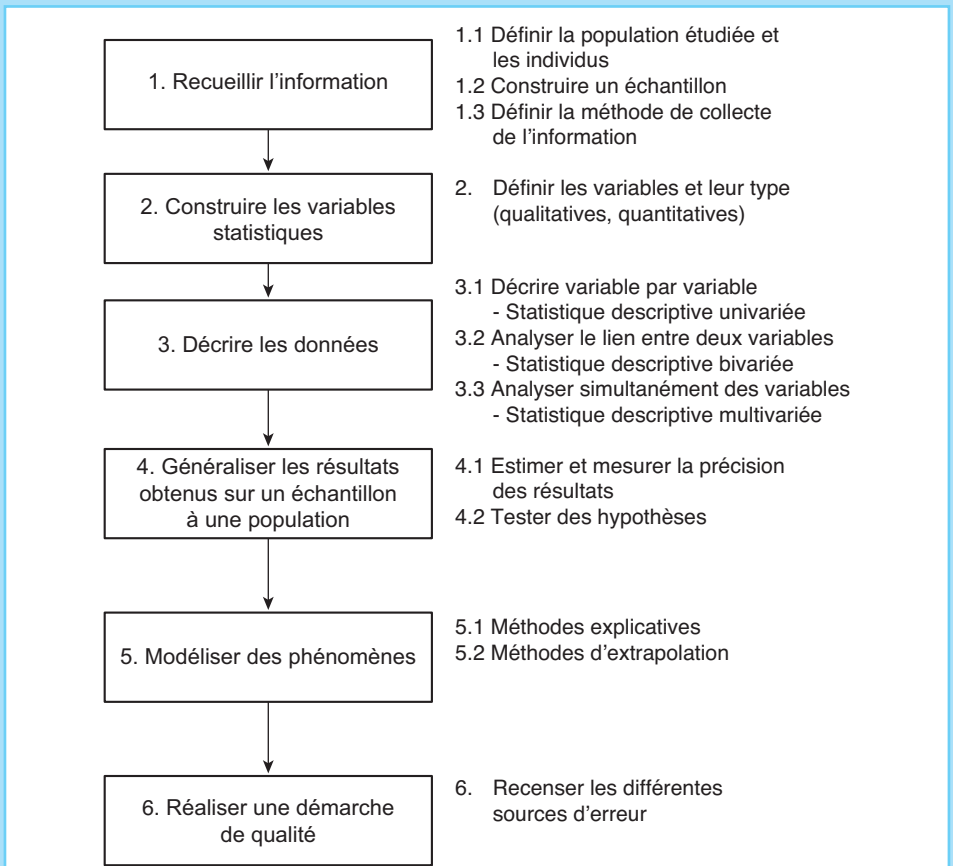
L'erreur totale dans les études statistiques.

Il faut également éviter qu'une action visant à diminuer un certain type d'erreur ne se reporte sur un autre. Par exemple, s'il est nécessaire de former et d'encadrer des enquêteurs, augmenter la taille d'échantillon a un impact direct sur l'erreur de mesure, rendant plus difficile une homogénéité du travail des enquêteurs.

Comment présenter une étude ?

Toute présentation managériale (orale ou écrite) doit comporter une partie consacrée à la méthodologie de l'étude, c'est-à-dire littéralement à une « discussion de la méthode ». La méthodologie récapitule la problématique, la population considérée dans le cadre de l'étude, la méthode et les outils de recueil de l'information. L'idéal est de décrire la méthodologie dès la première page de présentation. Le lecteur est ainsi capable d'apprécier les conclusions et leur robustesse.

Procédure générale



La démarche statistique appliquée au management en quelques points

La démarche statistique comporte plusieurs étapes :

- Recueil de l'information
  - définition de la population à étudier
  - construction d'un échantillon par une méthode aléatoire ou empirique
  - détermination du mode de recueil de l'information
- Détermination des variables statistiques et de leur nature
  - qualitatives : nominales ou ordinales
  - quantitatives : discrètes ou continues
- Analyse descriptive de l'information
  - analyse variable par variable : statistique univariée
  - analyse du lien entre deux variables : statistique bivariée
  - analyse simultanée des variables : statistique multivariée
- Généralisation des résultats à la population lorsqu'ils ont été obtenus à partir d'un échantillon
- Modélisation et quantification de la relation entre les variables
  - méthodes explicatives
  - méthodes d'extrapolation

Les erreurs dans une étude statistique :

- Il est nécessaire de les identifier pour les contrôler et les réduire.
- Elles sont présentes à chacune des étapes de réalisation de l'étude : conception, outil de mesure, terrain, analyse.
- Elles proviennent de l'outil de mesure développé, des individus interrogés, de l'analyste.
- L'erreur totale combine les différentes erreurs : erreurs de non-observation, erreurs de couverture, erreurs de mesure, erreurs d'échantillonnage, erreurs de modélisation.

Miniquiz

1. L'erreur d'échantillonnage reflète le fait que l'échantillon a été mal fait.

VRAI FAUX

2. L'erreur de non-réponse a peu d'impact sur la qualité des résultats.

VRAI FAUX

3. Il est possible de calculer la précision d'un sondage par la méthode des quotas.

VRAI FAUX

4. On relève la date d'émission de pièces de monnaie. Il s'agit d'une variable quantitative.

VRAI FAUX

5. Pour réaliser un sondage aléatoire stratifié, il est nécessaire de disposer de la liste des individus qui composent la base de sondage.

VRAI FAUX

Exercices

Exercice 1 : Identifier le type de variable

Les variables suivantes sont-elles quantitatives discrètes ou continues ? Qualitatives nominales ou ordinales ? :

- les calories dans un sandwich de fast-food ;
- le montant des droits d'inscription d'un diplôme supérieur ;
- le candidat pour lequel un électeur a voté aux dernières élections présidentielles ;
- la date de création d'un fonds de placement ;
- l'âge d'un fonds de placement ;
- le taux d'un crédit immobilier.

Exercice 2 : Identifier l'univers d'analyse (population), un individu statistique et le type de variable

1. Un auditeur s'intéresse au taux de factures erronées. Définir l'univers d'analyse, un individu statistique et la variable à étudier. Quelle est sa nature ?

2. Pour calculer la rentabilité d'une campagne d'e-mailing, on calcule plusieurs indicateurs : taux d'ouverture de l'e-mailing, taux de commandes, et montant moyen des commandes. Définir l'univers d'analyse, un individu statistique, les variables analysées et leur nature.

Exercice 3 : Identifier la méthode explicative à pratiquer

Pour les problématiques managériales ci-dessous, indiquer la méthode explicative que vous proposeriez, la variable à expliquer et les variables explicatives. Vous pouvez vous aider du tableau 1.7.

- Comment expliquer le *turn-over* et le pourcentage de salariés susceptibles de quitter l'entreprise ? L'ancienneté dans le poste, le salaire, la fonction exercée jouent-ils un rôle ?
- La rentabilité d'un titre dépend-il fortement ou faiblement de la rentabilité du marché ?
- Comment prévoir si une entreprise ou un particulier qui sollicite un emprunt remboursera et honorera ses échéances ?
- Pour un restaurant donné, la satisfaction globale exprimée par le client dépend-elle du jour de la semaine ?
- Comment mesurer l'impact des montants des liquidités sur la faillite d'une entreprise (variable qualitative) ?

Web Les corrigés ainsi que des exercices complémentaires sont disponibles sur le site compagnon.