Chapitre I Concepts de base de la statistique

Les exemples de ce chapitre sont dans le classeur Concepts de base de la statistique.xlsm.

I.1 Objet et objets de la statistique

I.1.a Objet de la statistique

La statistique est l'ensemble des techniques ayant pour objet de *décrire*, numériquement et graphiquement les **populations**. Population, individus, variables données et statistiques sont les objets de la statistique. Ses résultats sont des statistiques. La statistique a pour origine la démographie.

I.1.b Population, individus

Les populations dont la statistique s'occupe aujourd'hui ne sont plus seulement humaines.

Tout ensemble d'éléments, à condition qu'ils soient nombreux, peut être objet de la statistique. Par exemple :

- Les *pièces* fabriquées par une machine.
- Les *véhicules* du parc d'un loueur.
- Les *rues* ou encore les *carrefours* d'une agglomération.
- Des *communications téléphoniques*, des *dossiers de crédit* etc. etc.
- Les *jours* d'une période quelconque, par exemple comme l'année.

Les éléments des populations, qu'on continue encore d'appeler *individus*, sont aussi appelés *unités statistiques*. Il est équivalent de définir la population ou les individus, puisque la population est l'ensemble des individus auxquels on s'intéresse.

Tout travail statistique commence par expliciter le plus précisément possible la population à laquelle il se réfère. Voir, à ce sujet, le § I.8, page 31, et, en particulier, le § I.8.a.

I.1.c Observations, mesures, variables

Chaque individu de la population fait alors l'objet d'*observations* ou de mesures. On parle de **variables** observées ou mesurées et, pour les résultats, de **valeurs** ou *données* observées.

Par exemple le poids (*variable*) des valises embarquées en soute (*individus*), le nombre (*variable*) des véhicules passés en un point d'un réseau routier pendant les 2460 minutes (*individus*) d'une journée, ou la vitesse (*variable*) des mêmes véhicules (*individus*) mesurée au même point.

On voit, dans ce dernier exemple que tout en continuant à s'intéresser au trafic automobile, on peut, selon le point de vue, très rapidement changer d'individus.

Les variables mesurées doivent être *objectives*. Ce n'est que rarement le cas dans les enquêtes d'opinion, surtout à propos de sujets sensibles comme la politique ou la sexualité.

I.1.d Le problème à résoudre

La statistique est un outil d'aide à la décision. On fait de la statistique à l'occasion d'un "problème", de gestion par exemple.

Le choix de la population et des variables dépend de la façon de poser le *problème* qui est à l'origine de la démarche statistique.

Voici quelques exemples.

α) Absentéisme

On désire étudier l'absentéisme dans une entreprise (c'est le problème). Pour y parvenir on peut choisir de compter, sur une période donnée :

- Le *nombre de jours ouvrés d'absence* (c'est la variable) de l'ensemble (c'est la population) des *salariés* (ce sont les individus).
- Le *nombre de salariés absents* (c'est la variable) de l'ensemble (c'est la population) des *jours* (ce sont les individus) de la période sur laquelle porte l'étude.
- La *durée*, le *motif* (deux variables) etc. des *absences* (ce sont les individus).

β) Fréquentation d'un site touristique

On désire augmenter la fréquentation d'un site touristique (c'est le problème). Que faire ?
De la publicité, par exemple. Mais dans ce cas, de quoi doit-on parler ? Quels moyens de communication doit-on choisir ? Quelle cible viser ?
On peut proposer de :

- Compter le *nombre de visiteurs* (c'est la variable) du site, sur un ensemble (c'est la population) de *jours ouvrés* (ce sont les individus).

- Mesurer leur profil socio-économique, c'est-à-dire les variables *nationalité*, *âge*, *catégorie socioprofessionnelle*, *sexe*, *domicile*, *habitudes de consommations* et *satisfaction* d'un ensemble (c'est la population) le plus vaste possible de *visiteurs* (ce sont les individus).

χ) Transports urbains

Le maire d'une ville se demande comment améliorer la vie des habitants dans sa ville, en particulier réduire les embouteillages, le bruit, la pollution ? (c'est le problème).
On peut lui proposer de mesurer sur un ensemble (ce seront les populations) le plus important possible de :

- *Déplacements* (ce sont les individus, supposés parfaitement définis), les variables *origine*, *destination*, *objet*, *horaire*, *durée*, *itinéraire* etc.
- *Emplacements de stationnement à intervalle de temps réguliers* (ce sont les individus) s'ils sont ou non occupés (c'est la variable "*occupation*").
- Tronçons de rues à intervalle de temps réguliers (ce sont les individus), le taux de places de stationnement disponibles libres.

δ) Contrôle des livraisons

Un responsable de production se demande si son fournisseur répond bien au cahier des charges (c'est le problème). Il va donc contrôler, à chaque livraison (ce sont les individus) les variables quantités livrées, délais de livraison, nombre de manquants, nombre de défectueux, pour les comparer aux spécifications du cahier des charges.

ε) Service à la clientèle

On convient de contrôler l'efficacité d'un guichet de service à la clientèle (c'est le problème) en comptant le nombre de clients servis (c'est la variable) périodiquement (à l'heure, à la minute, à la journée, ce sont les individus) ou, autre façon de traiter le problème, le temps passé au guichet (c'est la variable) par les clients (ce sont les individus).

ϕ) Études de marché

Des responsables techniques et commerciaux doivent décider de l'avenir d'une gamme de produits. Quels modèles améliorer, supprimer, créer (c'est le problème) ? On ne peut évidemment pas prendre de décision sans connaître l'état actuel et prévisionnel du marché de ces produits. On peut proposer de mesurer pour chaque produit :

- Le *montant des achats* (c'est la variable) par les *clients* pendant une période donnée, par exemple le dernier exercice fiscal (ce sont les individus), répartis en différentes catégories (ce sont d'autres variables mesurées sur les clients).
- Les *quantités vendues* (variable) des produits pendant la même période (les individus).

γ) Répartition des recettes

On doit répartir entre plusieurs transporteurs, qui exploitent tout un ensemble de lignes de bus ou de trains dans une région, ou de remontées mécaniques sur un domaine skiable etc.,

les recettes générées par la vente d'un abonnement unique commun (billet carte ou forfait) valable pour l'accès à l'ensemble des produits offerts : C'est le problème. On pourrait :

- Mesurer, sur les abonnés (ce seront les individus) des variables rendant compte de l'utilisation de chaque équipement.
- Mesurer, pour chaque équipement (ce seront les individus) des variables rendant compte de leur utilisation par les clients.

Conclusion :

- Tant qu'on n'a pas bien précisé le problème, la population qu'on choisit d'étudier, les variables mesurées sur les individus, on ne peut pas commencer à *faire de la statistique*.
- La statistique ne peut jamais étudier qu'*une* population à la fois, mais un problème peut s'envisager de plusieurs points de vue, donc plusieurs populations successivement.

I.1.e Difficultés rencontrées dans la définition de la population

- La définition précise de la population n'est pas toujours immédiate :
  - Les *salariés* d'une administration comprennent-ils les non titulaires, les vacataires, les contractuels, ceux qui sont *détachés* à plus ou moins long terme d'une ou vers une autre administration ?
  - Les *clients* d'une entreprise se limitent-ils à ceux qui lui ont déjà acheté quelque chose, ou à ceux susceptibles de lui acheter un ou plusieurs produits ? Le fait qu'on ne dispose évidemment pas du fichier des clients potentiels ne dispense pas de préciser qui fait partie ou pas de la population.
- La population n'est pas toujours entièrement accessible :
  - Une fois définie la population des clients potentiels d'une entreprise, comment faire des observations, donc de la statistique, sans fichier de cette population ?
  - Certaines professions sont regroupées sous le même code APE (activité principale de l'entreprise). Il sera alors difficile d'étudier celle qui intéresse le statisticien.
  - Les "essais destructifs" interdisent évidemment d'observer toute la population.

I.2 Variables (ou caractères) statistiques, valeurs

On appelle variable chaque information dont on dit indifféremment qu'elle est recueillie, mesurée ou observée sur chaque individu de la population. Chaque individu présente sa valeur de la variable ou du caractère observé. On parle de variable parce que l'information n'est pas la même d'un individu à l'autre. On classe les variables :

- D'abord selon leur type mathématique, variables, qualitatives ou numériques, et ces dernières en quantitatives et ordinales.
- Ensuite selon leur rôle dans l'étude statistique (variables d'intérêt ou explicatives).

La statistique consiste en effet à rechercher des justifications à la variabilité des observations d'un individu à l'autre, souvent dans un souci de prévision. Cela conduit à rechercher des liens entre différentes variables. Les variables d'intérêt sont aussi dites, un peu abusivement, variables *dépendantes*, et les variables explicatives *indépendantes*.

I.2.a Variable ou caractère quantitatif

Lorsqu'on peut faire correspondre, *dans le contexte de l'étude*, à deux ou plusieurs individus une valeur unique qui résulte d'une opération mathématique, comme l'addition etc. sur les valeurs des chacun des individus, on dit que la variable ou le caractère est *quantitatif*.

Exemples

- La population est un ensemble de ménages. La variable le *nombre d'enfants*. C'est bien un caractère quantitatif puisqu'un ménage à 2 enfants et un autre à 3 enfants, ont, ensemble, 2 + 3 = 5 enfants.
- La population est un ensemble de salariés. On mesure le nombre de jours d'absence sur une période précise. C'est bien un caractère quantitatif puisque si un salarié a été absent 1 jour et un autre 3, les deux salariés ont été absents 1 + 3 = 4 jours, du point de vue de la production dans l'entreprise.
- La population est une série de voyages des bus d'une ligne. On mesure le temps mis pour aller d'un bout à l'autre de la ligne. C'est une variable quantitative puisque si l'autobus a mis 1 heure 52 minutes et 34 secondes pour faire la ligne, et une autre fois 1 heure, 47 minutes et 28 secondes, les deux voyages ont pris ensemble, 3 heures 40 minutes et 2 secondes.
- La population est un ensemble des salariés. On mesure le revenu annuel du salarié. C'est bien une variable quantitative puisque si un salarié a un revenu annuel de 15 000 et un autre un revenu annuel de 20 000, les deux salariés ont ensemble un revenu annuel de 35 000.
- On branche en série plusieurs condensateurs électriques (chacun a sa *capacité*). La *capacité totale* de l'ensemble se calcule par une opération mathématique sur les capacités individuelles. La capacité est donc bien une variable quantitative. Mais l'opération n'est pas l'addition ! L'opération n'est d'ailleurs pas non plus la même si on branche les condensateurs en parallèle.
- La population est l'ensemble de passagers d'un vol d'une compagnie aérienne.
  - On mesure le poids de leurs bagages en soute. C'est une variable quantitative puisque le poids total des bagages de tous les passagers est bien la somme des poids de chacun de leurs bagages, qui affecte la consommation de carburant.
  - On mesure la taille des passagers. Ce n'est pas, du point de vue de la compagnie aérienne, une variable quantitative, il est absurde de dire qu'un passager de 1,8 m

et un autre de 1,75, font ensemble 2,55 m ! La taille est, ici, une variable ordinale (voir plus loin) numérique.

Conclusion : Il ne suffit pas que les valeurs observées soient numériques pour que la variable soit quantitative.

I.2.b Variable ou caractère qualitatif, ou nominal, modalités

Les différentes valeurs d'une variable qualitative sont aussi appelées des *modalités*. Excel parle de catégories.

Tout ce qui n'est pas quantitatif est qualitatif. On ne peut pas attribuer une valeur unique à deux ou plusieurs individus qui ont des valeurs différentes, même numériques.

Exemples

- Si Martin est célibataire et Durant est marié, on ne peut, sauf cas particulier, rien dire, du point de vue matrimonial, de Martin et Durant.
- Si Martin trouve que c'est "très bien", et Durant que c'est "plutôt mal", il n'est pas possible de leur attribuer une opinion commune.
- Le contre exemple ci-dessus aux variables quantitatives.

I.2.c Variable numérique

Une variable quantitative est forcément numérique, mais une variable numérique peut être qualitative. En effet, les modalités d'une variable qualitative peuvent parfaitement être *codées en numérique* (célibataire = 0, marié = 1, etc.), par nécessité informatique, par exemple, sans que cela autorise pour autant à considérer ces variables, ni à les traiter comme des variables quantitatives.

I.2.d Variable booléenne

Il n'y a que deux modalités, `VRAI` et `FAUX`. Exemple : de la variable "Produit périmé".

I.2.e Variable ordinale

Une variable qualitative est qualifiée d'ordinale quand on peut tout de même ranger par valeurs croissantes ou décroissantes les valeurs de différents individus.

Exemples et contre-exemples

- Si lundi il a fait 12° et mardi 13°, on ne peut pas dire que lundi et mardi il ait fait 25°. Il ne s'agit donc pas d'une variable quantitative, mais d'une variable ordinale, à échelle numérique. On peut quand même dire que mardi, il a fait plus chaud que lundi.

- Si Martin mesure 1,8 m. et Durant 1,75 m, Martin est plus grand que Durant. On peut même ajouter que Martin mesure 5 cm de plus que Durant, soit 2,86% de plus que lui.
- On peut ranger les appréciations d'un produit : "Très mauvais", "Mauvais", "Indifférent", "Bon", "Très bon", et on peut même numériser les appréciations par des notes de – 2 à + 2, en passant par zéro = Indifférent, sans rendre la variable opinion quantitative. C'est d'ailleurs à proscrire, car, "Très bon" n'a aucun rapport avec "Bon", ni dans l'absolu, ni d'une personne à l'autre. En effet, si Martin juge le produit X très efficace (codé 05), tandis que Durant le juge seulement efficace (codé 04), ce dernier ne le juge pas 20% moins efficace ! Par contre, on peut dire que Martin juge le produit X *plus* efficace que Durant.
- Si Martin est célibataire, et codé 1, et Durant est marié, et codé 2, ensemble ils ne sont pas, si la modalité "divorcé" est codée 3, divorcés. Il n'y a non plus aucune raison objective de ranger les statuts matrimoniaux dans un ordre plutôt que dans un autre.

I.2.f Variable d'intérêt

Ce sont les variables qui font l'objet principal de l'étude statistique.

I.2.g Variable explicative

Le problème consiste toujours à mettre en évidence, l'influence que peuvent avoir sur chacune des variables d'intérêt, une ou plusieurs autres variables, que l'opérateur peut contrôler, et qui lui serviront à stratifier la population. Les commerciaux parlent de segmenter la clientèle.

Exemple

L'âge, le sexe, le statut professionnel des salariés ont-ils une *influence* sur (expliquent-ils) l'absentéisme, problème auquel on s'intéresse ? Ces variables font toujours partie des questionnaires d'enquêtes socio-économiques, car on analyse l'influence qu'ils peuvent avoir sur les variables d'intérêt de l'enquête.

Les variables d'intérêt et explicatives peuvent évidemment être qualitatives, ordinales, ou quantitatives, discrètes ou continues.

On dit que variables explicatives sont **indépendantes**, et les variables d'intérêt **dépendantes**.

Comme on écrit habituellement une fonction par $\mathbf{Y} = f(\mathbf{X})$ en mathématiques, on désigne par **Y** la variable d'intérêt et par **X** la variable explicative.

I.3 Quelques confusions trouvées dans des copies

I.3.a Confusion population – variable

La définition de la population et de la variable n'est pas si simple qu'elle paraît à la lecture de ce qui précède. Supposons qu'on soit chargé d'une étude statistique sur l'absentéisme dans une entreprise. Des propositions erronées comme par exemple :

α) Pour la définition de la population

- "Nombre de salariés" : La population n'est jamais un nombre, mais un *ensemble*.
- "Ensemble des salariés absents" : Il vaut mieux s'intéresser à la population des *salariés*, et ajouter la variable *durée d'absence*. Les salariés absents seront ceux pour lesquels cette durée est positive.
- "Durée des absences" : Il ne peut s'agir que d'une variable, pas d'une population. Une durée se mesure sur une absence. La population doit donc être l'"Ensemble des déclarations d'absence".

  Comment se rendre compte d'une pareille erreur ?

  Tout simplement en se posant la question suivante, à propos de l'individu *Durée* :

  "*Qu'est que l'on peut mesurer d'une durée, par exemple "4 jours"* ?" Si elle est égale à 0, supérieure à 5 jours, etc. On comprend vite que ce n'est pas à cela qu'on s'intéresse.

β) Pour la définition de la variable

L'erreur revient toujours à une incohérence avec la population sur laquelle elle est supposée devoir être mesurée. La variable annoncée ne peut pas être mesurée sur les individus choisis. Il suffit parfois d'inverser les rôles pour que le contexte soit correct.

- La variable "nombre de salariés absents" ne peut pas se mesurer sur l'ensemble des salariés (la population). Impossible de mesurer un nombre de salariés sur un salarié ! Cette variable ne peut être mesurée que sur la population des jours ouvrés : "On prend un jour ouvré, et on compte le nombre de salariés absents ce jour-là".
- La variable "Nombre de salariés" ne peut pas se mesurer sur la population des "durées des absences". L'erreur vient ici de ce qu'on pense déjà au résultat de l'enquête, ce qu'on appellera "tableau de distribution", défini au Chapitre II suivant, au § II.1.a, page 37.
- La variable "présence des salariés", ne peut se mesurer ni sur la population des salariés, ni sur celle des jours ouvrés, ni sur tout autre population que sur le croisement de l'ensemble des salariés avec l'ensemble des jours ouvrés d'une période, et, en effet, "un salarié, un jour donné" (c'est l'individu), est (c'est la variable "présence", booléenne) présent ("présence" = `VRAI`), ou pas ("présence" = `FAUX`).