Les statistiques descriptives
Deuxième Partie
Introduction

Emmanuel LESUEUR

Sommaire

1 Définition

«La statistique a pour objet l'étude, à l'aide de traitements mathématiques, de nombreux faits correspondant à l'observation d'un phénomène, dans le but de rendre compte de la réalité, d'essayer de l'expliquer et d'aider à la prise de décision» (J.Hubler, 1996)

Attention à ne pas confondre «La statistique» et «Les statistiques».

Au singulier, la « *Statistique* » désigne la science qui permet d'obtenir et de traiter des données.

Au pluriel, « *Les statistiques* » sont un ensemble des méthodes scientifiques visant à collecter, à analyser et interpréter des données numériques.

Au singulier, elle désigne également une grandeur calculée à partir des observations recueillies (ex : moyenne d'âge des élèves d'une même classe, balance commerciale de la France, nombre de tweets émis quotidiennement etc..).

Au pluriel, c'est aussi un ensemble de données numériques observées (Population, élections, emploi, ventes…)

On parle de *statistique descriptive* lorsqu'il s'agit de collecter les données, de les représenter sous forme de tableaux et de graphiques les plus clairs possibles pour en déduire des hypothèses de travail. Des calculs mathématiques permettent d'analyser les données (Moyenne, Médiane, Quartile, Variance, Ecart-Type…)

Sur la base de ces hypothèses, il est possible de tirer des conclusions valables et de prendre des décisions raisonnables, on parle alors de *statistique inférentielle*. On utilisera des tests d'hypothèses (par exemple le test du Khi-Deux), des méthodes de comparaison ; on pourra également modéliser certains phénomènes, en utilisant, par exemple, la méthode de la régression linéaire.

2 Histoire de la statistique

Les statistiques (descriptives) sont nées de l'activité de recueil des données (population, armée, impôts, organisation des richesses) répondant aux besoins d'organisation et de gouvernement des grands empires (Babyloniens, Hittites, Assyriens, Sumériens…). Par exemple, les premiers recensements connus ont été réalisés vers 3000 ans avant notre ère en Sumérie (Irak actuelle). Les techniques étaient très rudimentaires et leur mise en œuvre restait l'apanage du pouvoir politique. Les recensements de population et de ressources, les statistiques (du latin « status » c'est-à-dire État) sont restées purement descriptives jusqu'au XVII^{ème} siècle.

Le terme « statistique » est issu du latin « statisticum », c'est-à-dire qui a trait à l'État. Il a été utilisé, semble-t-il pour la première fois, à l'époque de Colbert, par Claude Bouchu, intendant de Bourgogne, dans une « Déclaration des biens, charges, dettes et statistiques des communautés de la généralité de Bourgogne de 1666 à 1669 ».

Puis s'est développé le calcul des probabilités et des méthodes statistiques sont apparues en Allemagne, en Angleterre et en France. Beaucoup de scientifiques de tous ordres ont apporté leur contribution au développement de cette science : PASCAL, HUYGENS, MOIVRE, BERNOULLI, LAPLACE, GAUSS, MENDEL, PEARSON, FISCHER etc....

Depuis les années 1960, avec le développement et la banalisation des outils informatiques et graphiques, la statistique, et surtout la statistique descriptive multidimensionnelle, a connu une expansion considérable.

Les statistiques sont aujourd'hui utilisées dans tous les secteurs d'activité :

- Industrie : fiabilité, contrôle qualité, ….
- Economie et finance : sondages, enquête d'opinion, assurance, marketing, …

- Santé
- Environnement, Météorologie
- Agronomie
- Sociologie

3 Vocabulaire et terminologie

Les statistiques descriptives visent à étudier les caractéristiques d'un ensemble d'observations comme les mesures obtenues lors d'une expérience. L'expérience est l'étape préliminaire à toute étude statistique. Il s'agit de prendre "contact" avec les observations. De manière générale, la méthode statistique est basée sur le concept suivant.

3.1 L'épreuve statistique.

C'est une expérience que l'on provoque. Par exemple, pour évaluer la durée de vie d'un filament on va faire brûler des ampoules jusqu'à extinction en mesurant le temps correspondant.

3.2 Population

C'est l'ensemble concerné par une étude statistique. On parle aussi de champ de l'étude. Par exemple, si l'on s'intéresse aux notes d'une promotion estudiantine, ce groupe constitue la population. A noter que si l'on s'intéresse maintenant à la production de voitures électriques, la population est alors constituée de **l'ensemble** des véhicules électriques fabriqués pendant une période donnée. Le terme de population et donc plus large en statistique que dans le langage courant. Cet ensemble est en général noté Ω.

3.3 Individu

On désigne ainsi tout élément de la population considérée. On parle aussi d'unité statistique. Dans les exemples indiqués ci-dessus, un individu est tout étudiant du groupe dans le premier cas et tout véhicule électrique fabriqué dans le second cas. Il est noté ω.

3.4 Échantillon

C'est un ensemble d'individus prélevés dans une population déterminée. Pour les exemples précédents, on ne prendra que les notes obtenues par les femmes. Pour le cas des voitures, on ne s'intéressera qu'aux véhicules à 3 portes.

3.5 Taille de l'échantillon

C'est le nombre d'individus dans l'échantillon. On parle de cardinal de l'échantillon, noté Card(Ω) ou n. On parle aussi d'effectif.

Si l'échantillon est constitué de **tous** les individus de la population, on dit que l'on fait un **recensement**. Il est extrêmement rare que l'on se trouve dans cette situation, essentiellement pour des raisons de coût. Quand l'échantillon n'est qu'**une partie** de la population, on parle de **sondage**. Le principe des sondages est d'étendre à l'ensemble de la population les enseignements tirés de l'étude de l'échantillon. Pour que cela ait un sens, il faut que l'échantillon soit représentatif de la population. Pour y parvenir, il existe des méthodes dites d'échantillonnage, dont nous ne parlerons pas ici.

3.6 Caractère ou Variable statistique

Des observations concernant un thème particulier ont été effectuées sur des individus. Les caractéristiques que l'on mesure ou observe forme ce que l'on appelle les **variables statistiques**. Par exemple, les notes des étudiants à l'examen de Statistique, les mentions qu'ils ont obtenues à leur

bac, leur genre, la couleur de leurs yeux, le Chiffre d'Affaires par PME, le nombre d'enfants par ménage, etc…On la note généralement VS.

Il existe différents types de variables statistiques :

Variable qualitative: C'est une caractéristique non numérique (genre, nationalité, CSP…). Elle s'exprime par des valeurs sur lesquelles les opérations arithmétiques de base (somme, moyenne,…) n'ont pas de sens. Si les valeurs peuvent être classées ou hiérarchisées (Satisfait, Assez Satisfait, Peu Satisfait, Pas Satisfait,..), la variable statistique sera dite **ordinale**. Dans le cas contraire, elle sera dite **nominale** (Ex les pays qu'on ne peut pas classer).

Variable quantitative : C'est une caractéristique numérique (âge, taille, salaire…). Elle s'exprime par des nombres réels sur lesquelles les opérations arithmétiques de base (somme, moyenne,…) ont un sens. Une variable quantitative peut être **discrète** (ou discontinue) lorsqu'elle est représentée par un nombre de valeurs entières, fini ou dénombrable (âge, nombre d'enfants par ménage…). Au contraire, elle est dite **continue** lorsque toutes les valeurs réelles sont susceptibles d'être prises (taille, salaire…).

Les variables quantitatives continues peuvent être regroupées en **classe**. Par exemple, si la variable est le poids, un individu qui pèse 76 kg appartiendra à la classe [70-80[. Lorsque les données sont regroupées en classe, il faut définir les **extrémités de classe**, c'est-à-dire, la *borne inférieure* et la *borne supérieure* de chaque classe. Aussi, il est nécessaire de préciser si les valeurs des extrémités sont incluses ou non dans la classe. En effet, **tous les éléments de la population étudiée doivent se retrouver dans une et une seule classe**.

La notation d'une classe est : [70-80[. « [70 » signifie que la valeur 70 est incluse dans la classe alors que « 80[» signifie que la valeur 80 est exclue de la classe. Pour des raisons pratiques, on retiendra comme extrémités de classes, des valeurs arrondies. Ainsi, on facilite les calculs de **l'amplitude** de la classe et son **centre**.

L'amplitude d'une classe est la différence entre la valeur de la borne supérieure et celle de la borne inférieure. Dans notre exemple, elle sera de (80-70), soit 10.

Le centre d'une classe est la moyenne des extrémités de la classe. Dans notre cas, il sera de $\frac{80+70}{2}$, soit 75.

Chaque classe est donc caractérisée par :

- ☞ Sa borne inférieure
- ☞ Sa borne supérieure
- ☞ Son amplitude
- ☞ Son centre

Une valeur que peut prendre un caractère s'appelle une **modalité**. Par exemple, si la caractéristique étudiée est la couleur des yeux, les modalités seront « Bleu », « Vert » et « Marron ».