La démarche à suivre aux statistiques

Les statistiques suivent quatre étapes principales :

Planification: En se basant sur les connaissances actuelles du problème, on décide quelles données doivent être prélevées et comment ?

Action: Nous recueillons les données en réalisant une expérience, en menant une étude, . . .

Analyse: Nous analysons les données afin de voir si nous avons obtenu une réponse à notre question initiale.

Prédiction: Nous mettons à jour nos connaissances et les utilisons pour savoir à quoi nos données futures pourraient ressembler.

Les statistiques comprennent:

- la collecte des données ;
- le traitement des données collectées, aussi appelé la statistique descriptive ;
- l'interprétation des données, aussi appelée l'inférence statistique, qui s'appuie sur la statistique mathématique.
- la présentation afin de rendre les données compréhensibles par tous.

1. Recueil des données

L'enquête statistique est toujours précédée d'une phase où sont déterminés les différents caractères à étudier et le choix de la population à étudier. Il se pose alors le problème de l'échantillonnage : choix de l'ensemble des individus à étudier (la population d'étudie), la taille de la population et sa représentativité.

De la situation de recherche au tableau de codage

Les données sont rassemblées dans un tableau à double entrée comprenant autant de lignes que d'unités d'observation indépendantes (le plus souvent, ce sont des sujets); autant de colonnes que de "variables" ou de mesures caractérisant chacune des unités d'observation.

Le tableau ainsi constitué se nomme un tableau de codage des données

- Les lignes de cette table représentent les unités d'observation.

Les observations sont réalisées sur un nombre limité d'unités ou d'individus dans le but de tirer des conclusions applicables à une population entière. Ce nombre limité d'unités d'observation forme l'échantillon.

- Quant aux colonnes du tableau de codage, elles représentent les variables.

Une variable, c'est une quantité ou qualité susceptible de fluctuations ou encore une grandeur à laquelle on peut attribuer plusieurs valeurs différentes, plusieurs niveaux. Au minimum, une variable a toujours deux modalités ou deux fluctuations. Ainsi, la variable sexe possède deux modalités: fille et garçon.

Une variable, telle le degré de concentration, peut être estimée sur une échelle à cinq degrés; elle a alors cinq modalités: très faible, faible, moyen, fort, très fort

Les variables placées en colonne dans le tableau de codage seront accompagnées d'une légende de codage reprenant l'ensemble de l'information nécessaire pour pouvoir reconnaitre la signification précise de n'importe quel code dans n'importe quel endroit du tableau.

La légende mentionne le numéro de la variable :

Exemple: V10 signifie variable n: 10,

V_{10}: sexe: 1= homme, 2 = femme

2. Traitement des données

Le traitement et l'interprétation des données ne peuvent se faire que lorsque celles-ci ont été collectées. La statistique a des règles et des méthodes sur la collecte des données, pour que celles-ci puissent être correctement interprétées.

- Le **pré traitement**

Le pré traitement des données est extrêmement important, en effet, une transformation des données initiales (un passage au logarithme, par exemple), peuvent considérablement faciliter les traitements statistiques suivants. Pour pouvoir exploiter les données, le regroupement des données, le calcul des effectifs et la construction de graphiques permettent un premier résumé visuel du caractère statistique à étudier. Il est parfois nécessaire de faire un classement, un résumé et d'opérer une compression et une transformation de données.

En statistiques, on est en général en présence d'un grand nombre de valeurs. Or, si l'intégralité de ces valeurs forme l'information, il n'est pas aisé de manipuler plusieurs centaines voire des milliers de chiffres, ni d'en tirer des conclusions. Il faut donc calculer quelques valeurs qui vont permettre d'analyser les données : c'est le rôle des réductions statistiques. Celles-ci peuvent être extrêmement concises, réduites à un nombre : c'est le cas des valeurs centrales et des valeurs de dispersion. Certaines d'entre elles (comme la variance) sont élaborées pour permettre une exploitation plus théorique des données Critères de position et Critères de dispersion. On peut aussi chercher à comparer deux populations. On s'intéressera alors plus particulièrement à leurs critères de position, de dispersion ou à l'analyse de la variance.

3. Interprétation et analyse des données

On suppose qu'une partie de la variation des données est due au hasard. On formule des hypothèses sur la loi qui gère le phénomène général. L'étude de l'échantillon va alors valider ou non cette hypothèse : c'est ce qu'on appelle **les tests d'hypothèses**. Ces tests permettent de quantifier la probabilité avec laquelle des variables vérifient une propriété donnée.

On peut chercher à modéliser un phénomène. La modélisation statistique doit être différenciée de la modélisation physique :

- La modélisation physique cherche à construire un modèle *explicatif* d'un phénomène, qui est soutenu par une théorie plus générale décrivant *comment* les phénomènes ont lieu en exploitant le principe de causalité.
- Dans le cas de la modélisation statistique, le modèle va être construit à partir des données disponibles. Ce type de modélisation s'appelle aussi **modélisation empirique**.

Un modèle est avant tout un moyen de relier des **variables à expliquer** Y à des **variables explicatives** X, par une relation fonctionnelle : $\boldsymbol{Y = F(X)}$

Les modèles statistiques peuvent être regroupés en grandes familles (suivant la forme de la fonction F):

- les modèles linéaires ;
- les modèles non linéaires ;
- les modèles non paramétriques.