Statistique élémentaire

Résumé

Statistique, fouille ou Science des Données, les appellations changent le volume et la diversité des données explosent, les technologies se succèdent, les modèles et algorithmes se complexifient. L'estimation devient un apprentissage, la prévision remplace l'explication.

Le parcours pour devenir data scientist *est structuré en quatre parties :*

Retour *à l'introduction générale*

Saison 1 *(L3) Statistique élémentaire, descriptive* vs. *inférentielle.*

Saison 2 *(M1) Statistique Exploratoire multidimensionnelle et apprentissage non supervisé.*

Saison 3 *Apprentissage Statistique / Machine supervisé.*

Saison 4 *(M2) Technologies pour la Science des (grosses) Données.*

plus des réflexions sur : Statistique et Déontologie scientifique.

1 Définitions

Le travail du statisticien est d'abord un travail de communication avec des représentants d'autres disciplines ou d'autres métiers. Ceci nécessite beaucoup de rigueur et donc de précision dans l'emploi des mots et concepts lorsqu'il s'agit de traduire en phrases intelligibles des résultats numériques ou graphiques. En effet, de ces *interprétations* découleront des prises de décision.

1.1 Statistique, statistiques, statistique

Le mot *statistiques* avec un "s" est apparu au XVIIIème siècle pour désigner des quantités numériques : des *tables* ou *états*, issus de techniques de dénombrement et décrivant les ressources économiques (impôts...), la situation démographique (conscription...), d'un pays. La Statistique est une sous-discipline des Mathématiques qui s'est développée depuis la fin du XIXème siècle notamment à la suite des travaux de l'école anglaise (K. Pearson, W. Gosset (Student), R. Fisher, J. Neyman...). Une *statistique* est une quantité définie par rapport à un modèle (*i.e.* une statistique de test) permettant d'inférer sur son comportement dans une situation expérimentale donnée.

1.2 Statistique descriptive vs. inférentielle

De manière approximative, il est possible de classer les méthodes statistiques classiques en deux groupes : celui des méthodes descriptives, celui des méthodes inférentielles.

- La *Statistique descriptive* regroupe les méthodes dont l'objectif principal est la *description* des données étudiées ; cette description des données se fait à travers leur *présentation* (la plus synthétique possible), leur *représentation graphique*, et le calcul de *résumés numériques*. Dans cette optique, il n'est pas fait appel à des modèles probabilistes. On notera que les termes de statistique descriptive, *statistique exploratoire* et *analyse des données* sont plutôt synonymes.
- La *Statistique inférentielle* regroupe les méthodes dont l'objectif principal est de préciser un phénomène sur une population globale, à partir de son observation sur une partie restreinte de cette population, l'échantillon. Il s'agit donc d'induire (ou encore d'inférer) du particulier au général avec un objectif principalement *explicatif*. Ce passage ne peut se faire qu'aux moyens de modèles et d'hypothèses probabilistes. Les termes de statistique inférentielle, *statistique mathématique*, et *statistique inductive* sont eux aussi quasiment synonymes.

2 Vocabulaire de la Statistique

Population Ω (ou population statistique) : ensemble (au sens mathématique du terme) concerné par une étude statistique. On parle parfois de *champ de l'étude*.

Individu $\omega \in \Omega$ (ou *unité statistique*) : tout élément de la population.

Échantillon : sous–ensemble de la population sur lequel sont effectivement réalisées les observations.

Taille de l'échantillon n : cardinal du sous-ensemble correspondant.

Enquête (statistique) : opération consistant à observer (ou mesurer, ou

questionner...) l'ensemble des individus d'un échantillon.

Recensement : enquête dans laquelle l'échantillon observé est la population tout entière (enquête *exhaustive*).

Sondage : enquête dans laquelle l'échantillon observé est un sous–ensemble strict de la population (enquête *non exhaustive*).

Variable (statistique) : $\Omega \xrightarrow{X} \begin{cases} \mathcal{E} \text{ si qualitative} \\ \mathbb{R} \text{ si quantitative} \end{cases}$ caractéristique (âge, salaire, sexe, glycémie...), définie sur la population et observée sur l'échantillon ; mathématiquement, il s'agit d'une application définie sur l'échantillon. Si la variable est à valeurs dans $\mathbb{R}$ (ou une partie de $\mathbb{R}$, ou un ensemble de parties de $\mathbb{R}$), elle est dite *quantitative* (âge, salaire, taille...) ; sinon elle est dite *qualitative* (sexe, catégorie socioprofessionnelle...). Si les modalités d'une variables qualitatives sont ordonnées (*i.e.* tranches d'âge), elle est dite *qualitative ordinale* et sinon *qualitative nominale*.

Données (statistiques) : ensemble des individus observés (échantillon), des variables considérées, et des observations de ces variables sur ces individus. Elles sont en général présentées sous forme de *tableaux* (individus en lignes et variables en colonnes) et stockées dans un fichier informatique. Lorsqu'un tableau ne comporte que des nombres (valeurs des variables quantitatives ou codes associés aux variables qualitatives), il correspond à la notion mathématique de *matrice*.

3 Démarche du statisticien

Le crédo de l'enseignant de statistique consiste à répéter inlassablement : un statisticien (ou les compétences qu'il représente) doit être associé *préalablement* à une étude, des expérimentations, une enquête... De la qualité du recueil et de l'organisation des données dépendra bien évidemment la *pertinence* des résultats de l'analyse. Plusieurs questions sont préalables :

3.1 Expérimentation

- Quelle est la question biologique, sociologique, épidémiologique à laquelle je veux apporter une réponse ? En particulier, quel est l'objectif (descriptif, explicatif, prédictif ou une combinaison) ?
- Quelle est la population étudiée ?
- Comment planifier des expériences ou des recueils d'informations dans des bases pré-existantes ?
- Quels sont les échantillons ?
- Précision des conditions expérimentales
- Observations et mesures

3.2 Exploration pour un objectif descriptif

Cette étape est de toute façon un préalable à tout autre objectif. Les données recueillies sont elles de qualité suffisante ? Sont-elles bien exemptes de biais ou artefacts expérimentaux ? Leurs grandes structures (groupes, corrélations...) sont-elles en accord avec les connaissances acquises sur le sujet ?

- Valeurs manquantes, erronées ou atypiques,
- Modalités trop rares,
- Distributions "anormales",
- Incohérences, liaisons non linéaires,
- Transformations, imputation, codage...

N.B. Ne jamais oublié que la préparation des données *data munging* représente plus de 80% du temps de travail avec des conséquence immédiates sur la qualité des résultats finaux : *garbage in garbage out*.

3.3 Décision pour un objectif explicatif

Telle variable ou tel facteur a-t-il une influence sur la variable d'intérêt ? Le modèle théorique est-il en accord avec les résultats expérimentaux ?

- Explicitation de l'hypothèse statistique répondant à la question,
- Détermination du modèle statistique correspondant,
- Estimation des paramètres du modèle et calcul de la statistique de test,
- Prise de décision : rejet ou acceptation de l'hypothèse.

Toutes les méthodes et techniques utilisées nécessitent d'être illustrées sur des exemples simples ou "académiques", pour ne pas dire simplistes, afin d'en comprendre les fondements. Néanmoins, leur apprentissage effectif requiert leur utilisation effective sur des jeux de données en vraie grandeur, issus de différents domaines d'applications. Ce n'est qu'à cette condition que peuvent être appréhendées les difficultés de mise en œuvre, les limites, les stratégies d'interprétation mais aussi la grande efficacité de ces outils.

3.4 Objectif de prévision

Consulter la saison 3 sans pour autant sauter la saison 2.

4 Déroulement

L'apprentissage de ces concepts, notions, techniques est initiée dans deux tutoriels.

- Description, tests de comparaisons puis modélisation en vue de la prévision de la concentration en ozone sur des données élémentaires.
- Même chose pour l'analyse d'une cohorte familiale.

Ceux-ci renvoient vers les vignettes résumant les notions de Probabilités et Statistique classique d'un cours au niveau L3. Une connaissance correcte et opérationnelle de ces notions évite de pratiquer la Science des Données en se libérant de la deuxième devise des shadoks (figure 1). Ne pas essayer toutes les options dans tous les sens jusque ça marche car, quand ça marche, il n'est pas sûr que ce soit vraiment les résultats avec la validité recherchée.

- Statistique descriptive unidimensionnelle
- Statistique descriptive bidimensionnelle
- Modèles probabilistes et variables aléatoires
- Estimation statistique
- Tests statistiques
- Régression linéaire simple
- Introduction à l'analyse en composantes principales
- Introduction à la régression multiple

FIGURE 1 – Première devise Shadok