OBJECTIF

STAPS

Licence et Master

Statistique et traitement des données

Du recueil à l'interprétation

- L'essentiel à connaitre
- Douze problématiques traitées en profondeur

Léo **Gerville-Réache**

ellipses

Je me rends compte que je parle de la statistique depuis plusieurs pages et que je n'en ai toujours pas donné la définition ! Pour être honnête, je repoussais ce moment car il n'existe pas de définition aussi univoque que celle d'un nombre entier en arithmétique ou d'un carré en géométrie. Pour ma décharge, on parle, par exemple, de « sciences de la vie » sans être capable de définir précisément le « vivant ». En revanche, on est tous d'accord pour dire qu'un chat que l'on voit courir dans la rue est vivant (c'est moins évident s'il s'agit du chat dans une boîte de Schrödinger). Être capable de classer en vivant ou non-vivant certaines entités n'est déjà pas si mal, même si dans certains cas, ce n'est pas encore tranché.

La statistique est une science, pas de doute. Une science qui traite d'informations et de données, pas de doute. J'aime dire que la statistique a pour but de réduire l'information pour gagner en compréhension. Mais ce n'est qu'une petite partie de son pouvoir…

Dans un « vieux » dictionnaire des mathématiques aux éditions PUF, on trouve la définition suivante : *Branche des mathématiques appliquées en liaison avec le calcul des probabilités mais qui, à la différence avec ce dernier, est basé sur des observations d'événements réels à partir desquelles on cherche à établir des hypothèses plausibles en vue de prévisions concernant des circonstances analogues.* Il y a du vrai dans cette définition…

Dans le secondaire, vous avez été initiés aux statistiques, via des tableaux de données, des graphiques, des calculs de pourcentages, de moyennes et d'écarts-types. En probabilité, vous avez mélangé des boules dans des urnes, calculé des combinaisons, des arrangements, découvert la loi de Bernoulli, voire la loi normale. Essentiellement on vous a séparé les problématiques. Certains d'entre vous ont croisé l'articulation entre statistique et probabilité via par exemple le concept d'intervalle de confiance.

La statistique sans probabilité existe ! On la nomme souvent : statistique descriptive. C'est une partie noble de la statistique, c'est une partie essentielle de la statistique, c'est une partie incontournable et loin d'être toujours simple.

Je vous disais que la statistique avait pour but, entre autres, de réduire l'information pour gagner en compréhension. Choisir la juste réduction, celle qui permet une juste compréhension, c'est déjà démontrer une maîtrise statistique. « Choisir », c'est le mot juste ! Faire de la statistique, c'est choisir, choisir de façon aussi objective et pertinente que possible les méthodes, les calculs, les graphiques qui permettent une interprétation juste. Plus facile à dire qu'à faire !

Pour vous donner un exemple, rien de tel qu'un contre-exemple ! Lê, de la chaîne Youtube Science4All (https://www.youtube.com/watch?v=0NbyYOcIwAY), nous propose la situation suivante : Vous devez choisir entre deux joueurs de foot pour une sélection dans votre équipe et vous hésitez entre Jack et Joe. Vous décidez alors de regarder si « *l'équipe gagne plus souvent lorsque le joueur en question*

joue ». Vous constatez alors que l'équipe gagne dans 50 % des cas lorsque Jack joue mais dans 80 % des cas lorsque c'est Joe qui joue. Vous en concluez que Joe est meilleur (ou plus utile) que Jack et décidez de le sélectionner. Bien joué !

À moins qu'en regardant plus en détail les statistiques et en particulier les équipes adverses qui ont participé aux calculs des pourcentages de victoires, vous vous rendiez compte que Joe a essentiellement joué contre des équipes faibles et Jack contre des équipes fortes. Aie ! Il est bien possible que votre statistique initiale ne soit pas si pertinente que cela pour faire votre choix. Cette situation sur laquelle je reviendrai, tant elle est dangereuse, est connue sous le nom de *paradoxe de Simpson*. Ce paradoxe emblématique de la statistique est là pour nous alerter sur les dangers d'une interprétation et d'une utilisation peu réfléchie d'un chiffre, quel qu'il soit.

La statistique est une science puissante qui peut se révéler redoutablement efficace. Le contre-coût est qu'il est nécessaire de bien en comprendre les mécanismes et les limites afin d'éviter au maximum ses pièges.

Dans toute science, il existe des fondamentaux. Il s'agit des principaux concepts et théories qui la caractérise. Une science est toujours en mouvement, en développement et sujette à réflexion épistémique. Pour autant, dans une époque donnée, ici les années 2020, on peut s'autoriser à faire un point d'étape, un constat. La démarche statistique peut se classifier, s'ordonner. La donnée, la description, la comparaison, la modélisation, l'enquête et enfin l'expérimentation constituent une architecture (de mon point de vue) pertinente des enjeux universels de cette science.

Dans ces « fondamentaux », je vais tenter de vous sensibiliser aux principaux concepts et théories qui me semblent donc incontournables. Voici le programme :

La donnée

Vous l'avez compris, la statistique est une science qui se fonde sur des observations, des données. Souvent nous les subissons. C'est-à-dire qu'elles préexistent généralement et nous ne sommes pas acteurs de leur production (nous verrons dans les parties « enquête » et « expérimentation » comment « produire » de la donnée pertinente). Dans cette partie, j'insisterai sur les différents types de données et leurs utilités, les principes de construction d'un tableau de données à vocation statistique, la définition d'indices et finalement sur un concept fort : celui de FAIR data.

L'outil informatique

Cela fait maintenant plusieurs décennies que l'informatique a pénétré notre quotidien. Mais cela n'a pas toujours été le cas. La statistique a profité de ce remarquable développement. De très nombreux logiciels libres ou payants ont alors vu le jour. J'en ai retenu un, il s'appelle R, il est totalement gratuit

et particulièrement puissant. Mais comme un très grand nombre d'analyses statistiques sont réalisées avec un tableur, je reviendrai également sur Excel et certains utilitaires peu connus.

La description

La prise en main d'un tableau de données est une étape essentielle qu'il ne faut surtout pas négliger. Il m'arrive encore de faire l'erreur de passer vite cette étape et après quelques heures d'analyses sophistiquées, de me rendre compte qu'une valeur aberrante ou encore une distribution curieuse m'avait bêtement échappé. C'est pour cela que cette partie commencera par discuter du résumé statistique. Vous découvrirez également le concept de loi normale, de valeur aberrante et de corrélation.

La comparaison

Le concept sûrement le plus fondamental et le plus utilisé est celui de la comparaison. On parle de comparaison statistique, de comparabilité statistique. Le raisonnement par comparaison n'est pas qu'une question statistique. Comparer fait partie des universels en sciences comme dans notre quotidien. La comparaison statistique nécessite beaucoup de précaution. Je vous parlerai de précision, grâce au concept d'intervalle de confiance, de test statistique et de théorie de la décision, et finalement de la star (un peu contestée) des statistiques : la p-value.

La modélisation

En matière de statistique, tout est affaire de modélisation, même si celle-ci n'est pas formellement précisée, mise en équation. Ici, j'irai davantage dans le détail. L'un des buts de la modélisation est de l'utiliser pour réaliser des prévisions, qu'il s'agisse d'interpolation ou d'extrapolation. Je vous parlerai de modèles probabilistes par « opposition » aux modèles déterministes, développerai le plus simple d'entre tous (la régression linéaire simple) et évoquerai des horizons plus subtils.

L'enquête

L'enquête fait partie des deux modes de recueils de données que l'on distingue généralement en statistique. C'est l'outil privilégié de beaucoup de sciences, sciences humaines et sociales en particulier. Vous verrez que la pertinence statistique d'une enquête dépend en premier lieu de l'échantillonnage. Les concepts de représentativité de l'échantillon et de marge d'erreur seront au cœur de mon propos.

L'expérimentation

Enfin, je vous parlerai de la démarche expérimentale qui ne peut se passer de la statistique. Qui dit expérience dit variabilité et donc statistique. Vous découvrirez la rigueur et l'ingéniosité qu'il faut mettre en œuvre pour constituer des groupes statistiquement comparables permettant de mettre en évidence des relations de cause à effet. Je vous parlerai en détail du paradoxe de Simpson, de la clause Ceteris Paribus et de randomisation.

La donnée

J'ai déjà évoqué la question de la différence entre observer et supposer, entre données et hypothèses, entre induction et déduction. En statistique, la donnée est la matière première. Une fois recueillie, on pourra la transformer, l'exploiter…

La statistique peut traiter de différents types de structures de données allant des plus simples (le tableau individus/variables) aux plus sophistiquées (flux temporel de bases de données en liaison). Dans cet ouvrage, d'une manière ou d'une autre, l'ensemble des données pourra se résumer à un tableau individus/variables. Mais rassurez-vous ce format de données permet de traiter de beaucoup de sujets !

1. Le tableau Individus/Variables

Une population (ou univers statistique) est un ensemble d'éléments auxquels on s'intéresse et sur lequel on désire obtenir des informations par le biais de statistiques. Un individu (ou unité statistique) est un élément de cette population. Un caractère ou variable est une caractéristique d'un individu. La première démarche d'une étude statistique est de parfaitement définir les caractères, les individus et la population étudiée afin de poser clairement les problèmes, de trouver les procédures statistiques adéquates et de les interpréter sans ambiguïté.

Dans un tableau Individus/Variables, chaque ligne représente un individu statistique. Cela peut être une personne mais aussi une entreprise, une date, un ménage, un club… Bref, en statistique, un individu est une entité appartenant à un groupe qu'il convient de choisir et définir en fonction de votre étude. Les individus doivent être de même nature.

Jour	Heure	Date	Journée	Domicile	SD	SE	Extérieur
Ven	20:00	17/08/2012	1	Stade Toulousain	23	22	Castres Olympique
Sam	14:00	18/08/2012	1	Aviron Bayonnais	6	13	ASM Clermont
Sam	17:30	18/08/2012	1	Biarritz Olympique	35	10	Stade Montois
Sam	17:30	18/08/2012	1	Stade Français Paris	32	16	Montpellier HR
Sam	17:30	18/08/2012	1	U Bordeaux-Bègles	28	29	FC Grenoble
Sam	17:30	18/08/2012	1	SU Agen	20	24	Racing Métro 92
Sam	19:40	18/08/2012	1	USA Perpignan	15	21	RC Toulon
Ven	19:50	24/08/2012	2	U Bordeaux-Bègles	26	22	USA Perpignan
Sam	14:00	25/08/2012	2	Racing Métro 92	21	23	RC Toulon
Sam	17:30	25/08/2012	2	Castres Olympique	30	13	FC Grenoble
Sam	17:30	25/08/2012	2	Stade Toulousain	37	22	Stade Montois
Sam	17:30	25/08/2012	2	Aviron Bayonnais	24	11	Stade Français Paris
Sam	17:30	25/08/2012	2	Montpellier HR	13	8	ASM Clermont
Sam	19:40	25/08/2012	2	SU Agen	19	25	Biarritz Olympique

Par exemple, on ne peut pas mélanger des individus « date » et des individus « club » dans un même tableau. D'autre part, d'une manière ou d'une autre, les individus sont supposés indépendants les uns des autres. C'est souvent le cas mais parfois, il faudra redoubler d'astuce pour que cela soit raisonnablement le cas.

La première ligne du tableau contiendra toujours les noms des variables. Chaque colonne du tableau sera donc une variable.

Dans l'exemple, il s'agit de rencontres du Top 14 (nous reviendrons sur ces données pour faire de la prévision…). Il contient 8 variables (Jour, Heure, Date, Journée du championnat, Équipe jouant à domicile, Score de l'équipe à domicile, Score de l'équipe à l'extérieur et Équipe jouant à l'extérieur) de deux types différents ; des quantitatives et des qualitatives.

Les variables quantitatives sont des variables pour lesquelles la distance entre deux valeurs de la variable a un sens. L'âge en années d'une personne, le chiffre d'affaires en euros d'une société, le nombre de points marqués par une équipe sont des variables quantitatives. Le fait que cette variable soit continue ou discrète a une importance lorsqu'on voudra faire des hypothèses sur sa distribution.

Les variables qualitatives sont les variables qui ne possèdent pas la propriété de « distance ». La couleur des yeux, le type de médicaments, le club de rugby… sont des variables qualitatives pour lesquelles on ne peut pas déterminer de distance entre les modalités. Les variables pour lesquelles on pourra quand même établir une relation d'ordre, entre les modalités (bon, moyen, mauvais par exemple), seront dites ordinales, les autres seront simplement nominales.

Je suis sûr que vous serez en capacité de bien identifier le type de chacune des 8 variables du tableau en exemple…

2. Les indices

Souvent, les variables recueillies permettent d'en calculer de nouvelles. Ce sont alors de nouvelles colonnes dans votre tableau. Par exemple, si vous disposez d'un tableau avec le poids et la taille de personnes, vous aurez peut-être une utilité à calculer l'IMC (Indice de Masse Corporelle). Je suis sûr que vous connaissez cet indice !

La construction d'indices est une étape extrêmement importante et souvent nécessaire pour réaliser ensuite des analyses statistiques utilisant ces indices. Je vous en rappelle et/ou présente quelques-uns ici (nous en croiserons d'autres plus tard…).

La *différence absolue* est souvent le premier type d'opération que l'on réalise. Par exemple, vous avez mesuré lors d'une séance d'entraînement, le pouls au repos puis après un effort sur les sportifs dont vous avez la charge. Vous pouvez calculer la différence (le delta) entre ces deux valeurs. Mais la *différence relative* serait peut-être plus pertinente. En divisant le delta précédent par la valeur du pouls au repos, vous obtenez le pourcentage d'augmentation du pouls lié à cette séance. Mais vous connaissez peut-être la fréquence cardiaque maximale de vos sportifs ! Pourquoi ne pas en tenir compte dans votre calcul ? C'est à vous d'y réfléchir…

Un indice peut également utiliser des statistiques obtenues sur une partie ou l'ensemble des individus. Un indice bien connu (et très utile) est le *z-score*. Il s'agit de relativiser chaque valeur d'une variable quantitative en prenant en compte la moyenne et l'écart-type de la variable. Pour être précis, pour chaque individu « i », on calculera son z-score pour la variable « X » de la manière suivante :

$$Z_i = \frac{X_i - \overline{X}_n}{s_n}$$

Où, $\overline{X}_n$ est la moyenne des valeurs de la variable X et s_n est l'écart-type de cette même variable. Si vous aimez les calculs, vous pourrez vérifier que tout z-score a une moyenne de 0 et un écart-type de 1. Vous pourrez également vérifier que les z-scores, comme les pourcentages d'ailleurs, sont sans unité de mesure. Ça c'est fondamental ! Cela signifie que leur interprétation pourra

souvent être générique. Cela signifie également qu'il sera possible de combiner des z-scores issus de variables initialement mesurées dans des unités différentes. Bref, d'une certaine manière, on pourrait ajouter des choux et des carottes (expression vraisemblablement désuète). Nous verrons plus tard plusieurs statistiques sans unité, je pense au coefficient de corrélation, au coefficient de variation ou encore à la p-value.

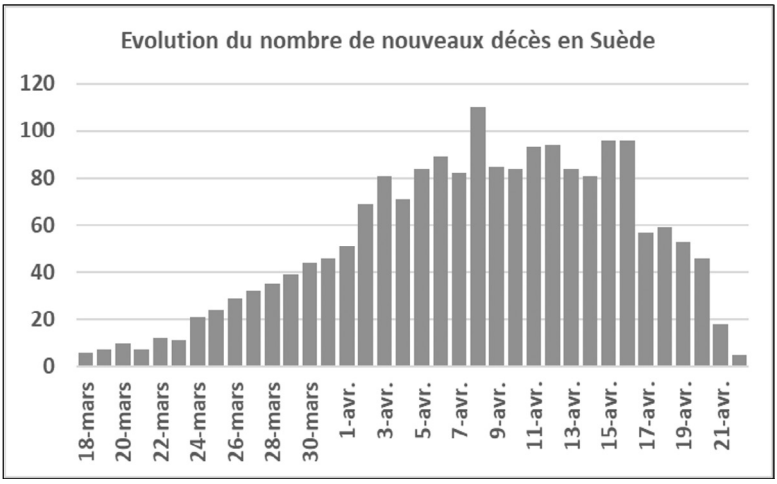
Un dernier indice pour la route… Il s'agit de la note. En effet, si vous avez mesuré par exemple des performances au saut en longueur sur un groupe d'étudiants en L1 STAPS, vous pouvez utiliser un barème de saut en longueur (adapté à ce public) et créer une nouvelle variable contenant les notes. Je reviendrai longuement sur la construction de barème dans la partie « problématique » de cet ouvrage.

3. Les FAIR Data

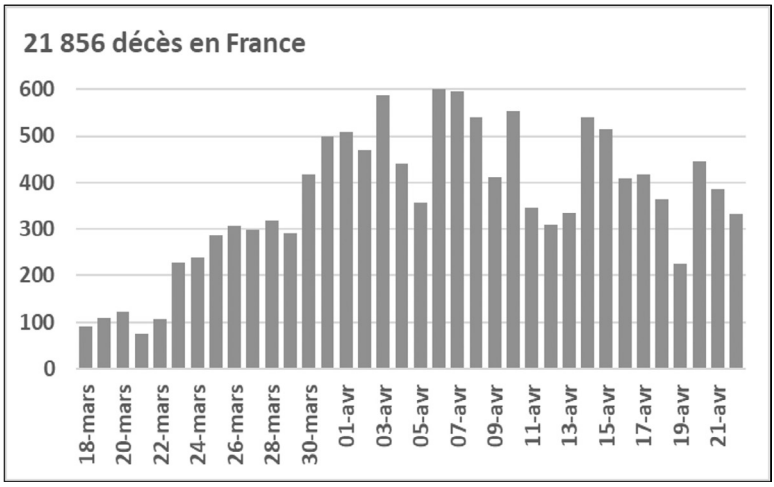
Pour introduire le concept de « FAIR data », je vous donne la définition que l'on peut trouver sur Wikipédia au moment où j'écris ces mots :

> *Dans le contexte de l'accessibilité de l'Internet, du big data (mégadonnées) des données de la recherche et des sciences ouvertes et plus largement du partage et l'ouverture des données, la notion de FAIR data ou données FAIR recouvre les manières de construire, stocker, présenter ou publier des données de manière à permettre que les données soient « faciles à trouver, accessibles, interopérables et réutilisables » 1 (en anglais :* findable, accessible, interoperable, reusable*), d'où l'acronyme « FAIR ».*

Je vous invite d'ailleurs à lire l'intégralité de l'article Wikipédia sur le sujet… Je vais ici me contenter de vous donner un exemple qui m'a donné du fil à retordre (encore une expression ancienne…). Nous sommes le 22 avril 2020, en plein 1er confinement (tout le monde sait de quoi je parle) et les médias (LCI ici) nous décrivent la situation en France mais également en Suède… En Suède, pas de confinement, les commerces et les restaurants sont ouverts… Et pourtant, le pic épidémique des décès est loin derrière eux ! Il aurait eu lieu autour du 10 avril. De plus, en 10 jours, le nombre de décès journaliers est passé de 100 par jour à presque 0 !



En même temps, en France, voici le type de graphique qui circule dans les médias. Pendant que la Suède avait 18 décès en 24 h le 21 avril, nous, en France nous en avions près de 400 ! Certes, nous avions passé le pic des décès mais c'est le confinement qui était interrogé. Où est l'astuce ?



Ah ! Les FAIR Data… Je vous explique : le 21 avril, en réalité, 185 décès dus au coronavirus ont été recensés en Suède, je dis bien 185 et pas 18. Allez voir sur le web les données officielles européennes, mondiales ou universitaires.

Mais alors ? En fait les Suédois produisent des statistiques de décès qu'ils actualisent tous les jours, les 185 décès du 21 avril sont ventilés sur des dates passées car ils ont visiblement des difficultés à compter les décès en temps réel. Voici ce que l'on peut produire sur 3 dates du mois d'avril :