1 -Vue d'ensemble sur les statistiques

Selon le Petit Larousse, la statistique est l'*« ensemble de méthodes mathématiques qui, à partir du recueil et de l'analyse de données réelles, permettent l'élaboration de modèles probabilistes autorisant les prévisions »*. Étymologiquement, le terme est dérivé du latin*Status*, l'État. Les statistiques sont ce qui est nécessaire pour gouverner un État.

Elles n'ont pas toujours bonne presse. Accusées de présenter des chiffres tendancieux ou d'ignorer les réalités humaines qui se vivent derrière, elles n'en sont pas moins un formidable moyen d'apporter des connaissances.

Le champ des statistiques

Le terme *statistiques* regroupe un ensemble de techniques que nous recensons ci-dessous (et de notations). Elles s'appuient sur les mathématiques. Mais les maths ne font pas « parler les chiffres », y compris dans leurs applications pratiques (recherche opérationnelle, mathématiques financières, physique...). Les problématiques sont différentes. Par ailleurs, la démarche mathématique est généralement déductive (on part d'une propriété générale pour démontrer une propriété particulière) au contraire de celle des statistiques, souventinductive (on extrapole les paramètres d'un échantillon à une population).

Survolons les diverses méthodes selon deux distinctions possibles.

La première d'entre elles sépare les techniques univariées des multivariées. Une technique univariée s'attache à une seule série d'un caractère donné ou à une seule mesure (même s'il y a plusieurs échantillons). Une technique multivariée analyse les éventuelles relations existant entre plusieurs caractères. Lorsque ceux-ci ne sont que deux, on parle d'analyse bivariée.

La seconde distinction est triple. D'abord, la technique peut être descriptive ; elle résume alors un ensemble d'observations, mettant en relief ce qui n'est pas ou peu perceptible sur le terrain. Ce sont les statistiques telles qu'elles sont perçues par le grand public depuis les années 60 (du moins en France), époque à laquelle les journaux télévisés se sont emparés des données chiffrées. La problématique peut être prédictive (ou inférentielle) auquel cas on établit un modèle probabilisé généralisable. Les prédictions s'appuient toujours sur des statistiques descriptives qui ne constituent souvent qu'une première étape de la démarche statistique. Enfin, la problématique peut être prévisionnelle. Elle s'appuie alors sur des techniques prédictives particulières, adaptées aux séries temporelles.

2- Univariées et descriptives

Une simple série statistique avec d'éventuels regroupements pour une présentation sous forme de tableau ne correspond qu'au sens vulgaire du mot *statistiques*. Pour mériter le label, il faut au moins calculer quelques informations synthétiques ! Ainsi, sur un seul caractère quantitatif, on peut établir la moyenne, l'écart-type, les quantiles… Quelques-unes de ces grandeurs sont enseignées au lycée. Vous trouverez tous les détails en pagedistribution univariée et en suivant les liens qui y figurent.

Les séries chronologiques peuvent être comparées entre elles lorsqu'on les traduit enindices simples. C'est une simple technique comparative, donc descriptive.

Ces types d'analyse sont ceux qui offrent le plus de possibilités de représentations graphiques. Si le caractère est qualitatif, on représente les proportions observées des différentes modalités (graphique circulaire, par exemple) sans donner lieu à des calculs.

3- Multivariées et descriptives

Quelques techniques bivariées permettent d'estimer si un lien existe entre deux variables quantitatives (corrélation) ou qualitatives (test d'indépendance du khi²).

Les techniques multivariées sont plus souvent nommées « analyses de données ». On connaît les valeurs prises par plusieurs caractères et l'on souhaite s'en servir soit pour déceler des proximités entre unités statistiques soit pour faire apparaître des groupes homogènes. Ou au contraire, on s'appuie sur des ressemblances entre unités statistiquespour montrer des proximités de caractères. Ce sont les techniques de classification qui sont alors utilisées. Un autre ensemble de techniques, les analyses factorielles, visent non seulement à trouver les proximités entre caractères et/ou individus mais aussi à déterminer les critères qui contribuent le mieux à « expliquer » les différences. Il en existe plusieurs. Certaines sont adaptées aux caractères quantitatifs, d'autres aux caractères qualitatifs.

L'analyse discriminante descriptive tient une place particulière dans la mesure où il s'agit d'une analyse factorielle dont le but est proche de celui d'une classification.

Les graphiques utilisés sont les nuages de points dans les plans factoriels et, pour un certain type de classification (en l'occurrence la CAH), les dendrogrammes. Une description bivariée de variables qualitatives est réalisable par stéréogramme.

Enfin, un mot sur les indices composites : on peut les qualifier de DESCRIPTIFS car, bien qu'établis sur des séries temporelles ils n'ont pas de finalité prédictive et de BIVARIÉS dans la mesure où ils font intervenir des prix et des quantités.

4- Univariées et prédictives

Considérons à présent un échantillon aléatoire dont on voudrait extrapoler quelques uns de ses paramètres (moyenne, proportion, variance) à une population totale, ou encore les comparer aux paramètres d'un ou plusieurs autres échantillons.

Ces paramètres auraient pu être différents car il existe des fluctuations d'échantillonnage. Ce sont donc des valeurs prises par une variable aléatoire.

Nous voici au pays des estimateurs et des tests.

En effet, on ESTIMERA le paramètre réel de la population à partir d'un paramètre observé sur l'échantillon.

Un test permettra d'accepter ou non une hypothèse sur un paramètre avec un risque d'erreur assumé. L'évaluation d'un risque repose évidemment sur des probabilités. Nous nous situons dans le cadre des statistiques probabilistes. Si la démarche vise à étendre à toute une population ce qui est observé sur un échantillon, on parle de statistiquesinférentielles.

Il faut connaître la loi de probabilité que suit le caractère observé. Certaines lois théoriques sont bien connues (loi normale, loi de Poisson…) et il est pratique de les utiliser parce qu'on peut alors employer

des tests dit « paramétriques » particulièrement efficaces. Encore faut-il pouvoir rattacher une distribution observée à l'une de ces distributions théoriques. Afin d'estimer si elle suit une loi en particulier, un premier test peut être réalisé (test d'adéquation à une loi ; voir tests de normalité, test de Kolmogorov, test d'adéquation du khi²). Si la distribution ne peut être rattachée à une loi théorique ou si les observations sont trop peu nombreuses, on utilise des tests non paramétriques, souvent moins puissants…

Les types de tests sont nombreux, certains étant applicables aux caractères quantitatifs et d'autres aux qualitatifs.

Ils ne donnent pas lieu à des représentations graphiques, la problématique étant plutôt de savoir si les graphiques de statistiques descriptives sur un échantillon peuvent représenter d'autres échantillons ou une population entière…

5- Multivariées et prédictives

Lorsque les caractères sont quantitatifs, la technique bivariée prédictive la plus connue est la régression linéaire simple. On s'intéresse à la relation entre une variable explicative et une variable dite expliquée. Il existe d'autres types de régressions simples, qui ne sont pas linéaires. La technique multivariée est la régression multiple. Les régressions peuvent être considérées comme simplement descriptives mais généralement, elles intègrent des notions probabilistes pour évaluer la qualité de leurs paramètres, ceci afin d'estimer leur capacité à prévoir.

Lorsque les caractères sont qualitatifs, les techniques sont l'ANOVA et l'ANOVA multivariée.

Mentionnons enfin l'analyse discriminante prédictive.

6- Prévisionnelles

La distinction entre univariées et multivariées n'est pas habituelle dans le cadre des techniques prévisionnelles, la plupart d'entre elles étant univariées.

Certaines font cependant du multivarié avec de l'univarié ! En effet, elles considèrent chaque observation comme une variable aléatoire particulière. Les diverses techniques figurent en page prévision des ventes. Si une régression multiple intègre une ou plusieurs variables « temps » parmi ses variables explicatives pour extrapoler la variable expliquée dans le futur, on peut la considérer comme une technique multivariée prévisionnelle.

Les graphiques associés sont toujours des courbes, éventuellement accompagnées d'un nuage de points…

7- Initiation aux séries statistiques discrètes

Remarque préalable : au sens strict, une série statistique est une suite de données individuelles. Donc, dès que des données sont regroupées sous forme de tableau, il ne s'agit plus d'une série statistique. Toutefois, cette page a été rédigée à l'attention des élèves de seconde pour qui cette distinction n'est pas de première importance. C'est pourquoi les deux situations seront vues ici.

Définitions

Une étude statistique porte sur des individus dont l'ensemble constitue la population. Un individu statistique n'est pas toujours un être humain et le terme d'unité statistique est souvent mieux adapté. Ce peut être une note, un pays, un relevé de température, bref, n'importe quoi. Le caractère (encore appelé variable statistique) est ce qui est observé sur chaque unité.

Ce caractère peut être qualitatif, discret ou continu (le cas continu ne sera pas traité ci-dessous mais en page série statistique continue).

Un caractère QUALITATIF ne se mesure pas sur une échelle numérique. C'est par exemple une couleur ou une profession. Un caractère DISCRET se mesure par des nombres entiers (ou avec un nombre limité de décimales) : nombre d'enfants, de clients… Un caractère CONTINU peut prendre des valeurs intermédiaires (poids, prix…) mais on le traite parfois comme un caractère discret en arrondissant les vraies valeurs (l'âge, par exemple).

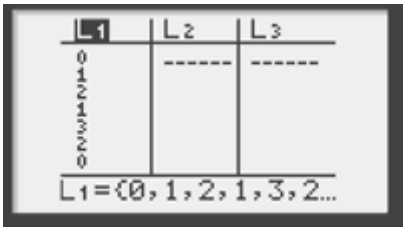
Avec calculatrice

À l'ère du *big data*, il faut souvent traiter de très nombreuses données, parfois des milliards. Bien entendu, elles n'existent que sous forme numérique. En classe de seconde, on traite un petit nombre de données avec une calculatrice. Nous verrons ici comment utiliser la TI-82 ou la TI-83. Pour un traitement manuel, voir la page exemple d'une série discrète. Avec une Casio : statistiques avec calculatrice Casio.

Dans un énoncé d'exercice, les valeurs d'une série sont séparées par des points-virgules. Par exemple, les nombres d'enfants d'une population de dix couples sont présentés ainsi : 0 ; 1 ; 2 ; 1 ; 3 ; 2 ; 0 ; 5 ; 2 ; 2.

La série peut être ordonnée. On le fait pour calculer à la main certains indicateurs (médiane, quartiles, fréquences cumulées) mais si le calcul est informatique (ordinateur ou calculatrice), le tri est inutile. En l'occurrence, la série triée est : 0 ; 0 ; 1 ; 1 ; 2 ; 2 ; 2 ; 2 ; 3 ; 5.

Sur la calculatrice TI-82, touche *stats* puis optez pour le choix 1 (EDIT). Dans la colonne qui se présente (L1), tapez chaque valeur suivie d'*entrer* (ou flèche vers le bas) pour aller à la ligne. Une erreur se corrige par la touche *suppr*. Une série entière s'efface en remontant sur la ligne d'en-tête (L1), puis touche *annul*, puis *entrer*. Les écrans de la TI-83 illustrent la page statistiques avec TI, que nous vous recommandons de lire...



Au cas où, voici comment trier ces valeurs dans l'ordre croissant : touche *stats* pour revenir au menu, puis, toujours dans le menu d'édition, choix 2 (TriA). Le choix 3 permettrait un tri par ordre décroissant. Lorsqu'on valide, la calculatrice affiche TriA(et l'on doit entrer le nom de la série. Ici, c'est L1. Attention, il ne faut pas taper la lettre L puis le chiffre 1 mais sur la touche *2nd*e puis sur la touche *1* (au-dessus de laquelle est imprimé *L1* de la même couleur que la touche *2nde*). On valide et la calculatrice nous indique que le travail est fait. Pour le vérifier, on peut revenir voir notre série (touche *stats*, choix 1, *Entrer*).

Un énoncé peut partir d'un tableau et non d'une série brute, surtout si la population est nombreuse. Dans la vie professionnelle également, on peut travailler soit à partir de données brutes, soit à partir de données regroupées. Les calculatrices gèrent aussi bien cette présentation que la précédente (ce qui n'est pas toujours le cas des logiciels).

Dans un tableau, on indique sur une première ligne les différentes valeurs que peut prendre le caractère et dans une seconde, le nombre de fois où cette valeur apparaît (en d'autres termes : l'effectif). Ici, nous aurions…

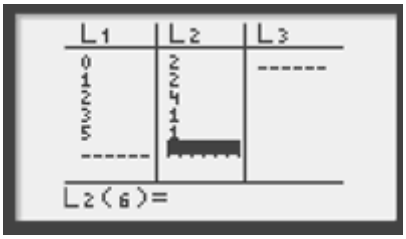
Nombre d'enfants	0	1	2	3	4	5
Nombre de couples	2	2	4	1	0	1

Précisons que la colonne qui correspond à quatre enfants est facultative puisqu'aucun couple de notre population n'a quatre enfants.

Nous vérifions que la somme des couples (seconde ligne) est bien égale à 10.

Dans les formules, les valeurs prises par le caractère sont notées xi.xi. **Le** ii **est un compteur. Par exemple, ici,** x1=0,x1=0, x2=1,x2=1, **etc. L'effectif qui correspond est noté** ni.ni. **Donc, ici,** n1=2,n1=2, n2=2,n2=2, **etc.**

Dans l'éditeur statistique de votre calculatrice, entrez dans la colonne L1 les valeurs prises par le caractère (nombre d'enfants) et dans la colonne L2 les effectifs correspondants.



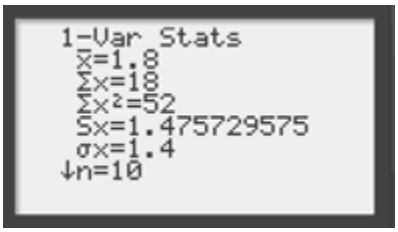
Maintenant, quelles informations peut-on obtenir ? Les différents indicateurs sont la moyenne, l'écart-type, la médiane, les quartiles, l'étendue… Ces notions sont largement traitées sur ce site web mais les pages qui les concernent ne sont pas toujours adaptées au niveau d'une classe de seconde. Pour la médiane, voir la page sur les fréquences.

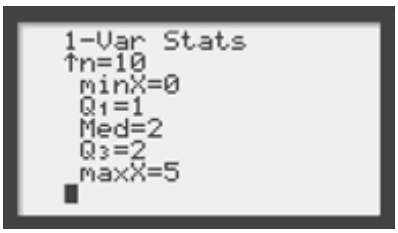
L'étendue est la différence entre la valeur la plus grande et la plus petite. Dans notre exemple, $5-0=5.5-0=5$.

La première information est la moyenne pondérée. Habituellement, les enseignants demandent que soit posée la formule, même si le calcul est effectivement réalisé avec la fonction STAT de la calculatrice. Dans notre exemple :

$\overline{x}=(0\times2)+(1\times2)+(2\times4)+(3\times1)+(5\times1)10=1,8\bar{x}=(0\times2)+(1\times2)+(2\times4)+(3\times1)+(5\times1)10=1,8$

Pour obtenir les indicateurs d'une série statistique avec la TI-82, il faut se rendre dans le menu CALC (touche *stats*). Optez pour le choix 1 qui correspond à une série d'une variable (les autres choix ne font pas partie du programme de seconde, ni même de première, d'ailleurs). Dans le premier cas où chaque valeur est entrée individuellement et si les données figurent bien en colonne L1 , il suffit d'appuyer deux fois sur la touche *Entrée* pour obtenir les écrans de résultats (il y a deux écrans à la suite, voir plus bas). Dans le second cas (L1 et L2), appuyez sur *Entrée* une seule fois puis *L1,L2* (la touche virgule n'est pas celle des décimales mais celle qui se situe au-dessus de la touche *7*). *Entrée*.





La première information est la moyenne (on retrouve bien 1,8), la suivante est la somme des xinixini **(c'est le numérateur de la formule), la troisième est la somme des carrés (inutilisée en seconde), les deux suivantes sont les écarts-types, n est l'effectif (très pratique pour s'assurer que toutes les données ont bien été entrées). Puis viennent le minimum, le premier quartile, la médiane, le troisième quartile et la valeur maximale. Attention, les définitions des quartiles Q1 et Q3 ne sont pas exactement les mêmes que celles du programme de seconde.**

La calculatrice peut aussi dessiner des graphiques mais cette fonctionnalité n'est pas d'un grand intérêt…

NB. Ce cite est très utilisable dans le domaine de formation SNV :

http://www.jybaudot.fr/Stats/stats.html