Statistique descriptive

1 Rappels

Définition 1 *La statistique (ou les statistiques) est une branche des mathématiques basée sur les observations d'événements réels à partir desquelles on cherche à établir des hypothèses plausibles en vue de prévisions concernant des circonstances analogues. L'étude d'un problème statistique peut se décomposer en quatre étapes : recueil de données, classement et réduction de ces données (statistique descriptive), analyse de ces données visant à la déduction de prévisions (statistique inférentielle).*

Nous allons nous contenter de faire de la statistique descriptive.

Une étude statistique descriptive s'effectue sur une **population** (des personnes, des villes, des voitures...) dont les éléments sont des **individus** et consiste à observer et étudier un même aspect sur chaque individu, nommé **caractère** (taille, nombre d'habitants, consommation...).
Il existe deux types de **caractère** :

1. **quantitatif** : c'est un caractère auquel on peut associer un nombre c'est-à-dire, pour simplifier, que l'on peut "mesurer". On distingue alors deux types de caractère quantitatif :
   - **discret** : c'est un caractère quantitatif qui ne prend qu'un nombre fini de **valeurs**. Par exemple le nombre d'enfants d'un couple.
   - **continu** : c'est un caractère quantitatif qui, théoriquement, peut prendre toutes les valeurs d'un intervalle de l'ensemble des nombres réels. Ses valeurs sont alors regroupées en **classes**. Par exemple la taille d'un individu, le nombre d'heures passées devant la télévision.
2. **qualitatif** : comme la profession, la couleur des yeux, la nationalité. Dans ce dernier cas, "nationalité française", "nationalité allemande" etc... sont les **modalités** du caractère.

En général une série statistique à caractère discret se présente sous la forme :

Valeurs	x_1	x_2	.....	x_p
Effectifs	n_1	n_2	.....	n_p
Fréquences	f_1	f_2	.....	f_p

Plutôt que réécrire ce tableau on écrira souvent : la série (x_i, n_i). (On n'indique pas le nombre de valeurs lorsqu'il n'y a pas d'ambigüité). Souvent on notera N l'effectif total de cette série donc $N = n_1 + n_2 + \ldots + n_p$.

Lorsqu'une série comporte un grand nombre de valeurs, on cherche à la résumer, si possible, à l'aide de quelques nombres significatifs appelés **paramètres**. En seconde vous avez défini les notions de moyenne (à rapprocher de la notion de barycentre...), médiane, mode qui sont des paramètres de position et la notion d'étendue qui est un paramètre de dispersion. Le but de ce cours est de définir de nouveau paramètres.
Dans la suite, tout caractère considéré est quantitatif.

2 Paramètres de position : les quartiles

Comme pour la médiane qui permet de partager l'effectif en deux effectifs égaux, intuitivement, les quartiles sont des nombres qui partagent la série statistique en quatre parties qui ont toutes "sensiblement" le même nombre de termes, c'est-à-dire 25% de l'effectif total.

2.1 Définitions

Définition 2
Le premier quartile Q_1 est la plus petite valeur du caractère telle qu'au moins 25% des termes de la série aient une valeur du caractère qui lui soit inférieure ou égale.
Le troisième quartile Q_1 est la plus petite valeur du caractère telle qu'au moins 75% des termes de la série aient une valeur du caractère qui lui soit inférieure ou égale.

Remarque :
Les définitions en statistique ne sont pas figées...certaines calculatrices et logiciels utilisent une définition différente ce qui explique que les résultats obtenus à l'aide de la "machine" soit différents de ceux donnés par la définition 2

2.2 Caractère discret

Dans ce cas, la définition 2 se traduit comme suit :
On commence par classer les valeurs x_i par ordre croissant, chacune d'elles répétées autant de fois, dans cette liste, que son effectif n_i, alors :

- Si $\frac{N}{4}$ est un entier, le premier quartile Q_1 est le terme qui dans cette liste occupe le rang $\frac{N}{4}$ et le troisième quartile est le terme de rang $\frac{3N}{4}$.
- Si $\frac{N}{4}$ n'est pas un entier, le premier quartile Q_1 est le terme de rang immédiatement supérieur à $\frac{N}{4}$ et le troisième quartile est le terme de rang immédiatement supérieur à $\frac{3N}{4}$.

2.3 Caractère continu

Dans ce cas on ne connaît pas chaque valeur du caractère il est donc imposible de mettre en place la définition. On se contente alors de valeurs approchées (sans connaître la précision ...) des quartiles. Pour cela différentes procédures sont possibles :

- On peut comme pour la médiane, tracer le polygone des fréquences cumulées croissantes et on "adopte" les valeurs suivantes :
  - Q_1 est la valeur correspondant à la fréquence cumulée croissante égale $0,25$.
  - Q_3 est la valeur correspondant à la fréquence cumulée croissante égale $0,75$.

  Quelques fois la lecture peut se faire sur la tableau des effectifs ou des fréquences cumulées croissantes...
- On peut aussi se contenter des classes contenant Q_1 et Q_3
- On peut, avec l'hypothèse que la répartition dans chaque classe est régulière, remplacer chaque classe par son centre pour se ramener à un cas discret.

2.4 Effet d'un changement affine

Théorème 1 *$(x_i; n_i)$ est une série statistique de médiane M_x, de quartiles Q_{1x} et Q_{3x}. La série de* **même effectif** *(y_i, n_i), telle que pour tout i, $y_i = ax_i + b$ $(a \in \mathbb{R}^*, b \in \mathbb{R})$ a :*

- *pour médiane $M'_y = aM_x + b$;*
- *pour quartiles, si $a > 0$, $Q_{1y} = aQ_{1x} + b$ et $Q_{3y} = aQ_{3x} + b$.*

Preuve

La preuve repose sur l'utilisation de la fonction $x \longmapsto ax + b$ qui est strictement croissante lorsque $a > 0$.

Remarque : Ce théorème peut-être utile lorsque l'on change le caractère d'unité (par exemple de francs en euros...).

3 Paramètres de dispersion

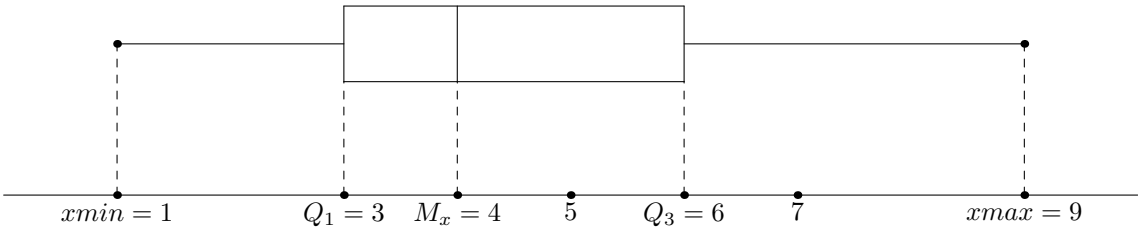
3.1 Écart inter-quartile

Définition 3 *L'* **intervalle interquartile** *est l'intervalle* $[Q_1; Q_3]$.
L' **écart interquartile** *est le nombre* $Q_3 - Q_1$. *C'est la longueur de l'intervalle interquartile.*

Remarque : Contrairement à l'étendue, l'écart interquartile élimine la valeurs extrêmes : ce peut être un avantage. En revanche il ne prend en compte que 50% de l'effectif : ce peut être un inconvénient.

3.2 Diagramme en boîtes

On construit un diagramme en boîte de la façon suivante :
- les valeurs du caractère sont représentées sur un axe (vertical ou horizontal) ;
- on place sur cet axe, le minimum, le maximum, les quartiles et la médiane de la série ;
- on construit alors un rectangle (c'est la fameuse boîte...) parallèlement à l'axe, dont la longueur est l'interquartile, la largeur est elle arbitraire.



Remarque : Ce diagramme permet non seulement de visualiser la dispersion d'une série mais aussi de comparer plusieurs séries entre elles.

3.3 Variance et écart-type

3.4 Introduction

Donnons-nous une série statistique quelconque à **caractère quantitatif discret** $(x_i; n_i)$. L'idée de cette section est de pouvoir "mesurer" la "dispersion" de l'ensemble des valeurs x_i de la série autour de sa moyenne $\bar{x}$. Pour cela on "mesure" pour chaque valeur x_i son "éloignement" par rapport à la moyenne puis on calcule "l'éloignement" moyen. Le tout est de décider comment "mesurer" pour chaque valeur son éloignement par rapport à la moyenne.

Exercice 1

On considère la série suivante :

Valeurs x_i	70	72	74	75	78	80	83
Effectifs n_i	2	1	2	2	1	3	1
Fréquences f_i							

1. Calculez la moyenne $\bar{x}$ de cette série.

2. Complétez le tableau suivant proposant trois façons de "mesurer" pour chaque valeur l'éloignement par rapport à $\bar{x}$.

$x_i - \bar{x}$							
$\lvert x_i - \bar{x}\rvert$							
$(x_i - \bar{x})^2$							

3. Calculez dans chacun des trois cas l'éloignement moyen. Conclusion ?

Pour une série quelconque, notons N l'effectif total :

- on appelle l'écart algébrique moyen le nombre :

$$l_m = \frac{1}{N}\sum_{i=1}^{p} n_i(x_i - \bar{x}).$$

  Ce nombre est toujours nul (preuve à étudier à titre d'exercice...) et ne permet pas de distinguer deux séries.
- on appelle l'écart absolu moyen le nombre :

$$e_m = \frac{1}{N}\sum_{i=1}^{p} n_i|x_i - \bar{x}|.$$

  Ce nombre fournit un très bon paramètre de dispersion mais il n'a pas d'application en statistique mathématique entre autres raisons parce que la valeur absolue se prête peu aux calculs. On s'intéresse alors à la moyenne pondérée des nombres $(x_i - \bar{x})^2$ qui a permis de formuler de nombreuses propriétés en statistique et en probabilité, vous le verrez au fur et à mesure de vos études.

3.5 Définitions et théorème

Définition 4
On appelle **variance** *d'une série quelconque à caractère quantitatif discret le nombre :*

$$V = \frac{1}{N}\sum_{i=1}^{p} n_i(x_i - \bar{x})^2 = \sum_{i=1}^{p} f_i(x_i - \bar{x})^2$$

L'écart-type *de cette série est* $s = \sqrt{V}$.
Si la série est regroupée en classes *ou si la caractère est quantitatif continu, avec l'hypothèse d'une* **répartition uniforme** *à l'intérieur de chaque classe, on remplace chaque classe par son centre. On est ainsi ramené à un cas discret.*

Remarque :

- On est amené à considérer la racine carrée de la variance pour avoir un résultat exprimé dans la même unité que le caractère étudié.
- Il existe un autre moyen de calculer V qui évite le calcul de $x_i - x$, le théorème suivant précise cette possibilité :

Théorème 2 *Théorème de Koenig (admis)*

$$V = \frac{1}{N}\sum_{i=1}^{n} n_i x_i^2 - \bar{x}^2$$

3.6 Propriétés de la variance

On a choisi de calculer la moyenne des carrées des écarts par rapport à la moyenne ; le théorème suivant donne une bonne raison de faire ce choix.

Théorème 3 *La fonction* $g : t \longmapsto \frac{1}{N}\sum_{i=1}^{p} n_i(x_i - t)^2$ *admet un minimum atteint en* $t = \bar{x}$ *(la moyenne de la série) et ce minimum vaut* V *(la variance de la série).*

Preuve
détaillée en cours, elle repose sur la dérivation de cette fonction g et un peu d'aisance technique.

Théorème 4 *$(x_i; n_i)$ est une série statistique de variance V_x, d'écart-type s_x. La série de* **même effectif** *(y_i, n_i), telle que pour tout i, $y_i = ax_i + b$ $(a \in \mathbb{R}^*, b \in \mathbb{R})$ a pour variance $V_y = a^2 V_x$ et pour écart-type $s_y = |a| s_x$.*

Preuve

Elle repose sur la fait que $\bar{y} = a\bar{x} + b$.

4 Résumés d'une série par ses paramètres

Le choix d'un résumé d'une série statistisque par ses paramètres n'est pas des compétences du mathématicien, ce sont celles des statisticiens, des économistes... suivant ce qu'ils veulent en faire. En tous cas, une étude statistique est accompagnéé de commentaires qui justifient la méthode employée et les choix faits. On peut cependant indiquer les résumés possibles suivants :

- Le couple (médiane ; étendue)
- Le couple (moyenne ; étendue)
  Ces deux couples sont simples à obtenir mais ils ne permettent pas de positionner le maximum et le minimum de la série. De plus l'étendue est un caractère de dispersion très grossier car sensible aux valeurs extrêmes.
- Le couple (médiane ; intervalle interquartile)
  Il est insensible aux valeurs extrêmes.
- L'ensemble {minimum, premier quartile, médiane, troisième quartile, maximum}.
  Il permet de construire un diagramme en boîte et donc de mieux visualiser le comportement d'une série (notamment sa dispersion) et de comparer des séries. Il présente un inconvénient : la connaisance de ces paramètres pour deux séries ne permet pas de calculer les paramètres du regroupement des deux séries.
- Enfin, le couple (moyenne, écart-type).
  Ce couple permet de faire des calculs sur des regroupements (cf exo...) et il permet à l'aide de l'inégalité de Bienaymé-Tchebychev (c'est pour plus tard...) d'avoir une idée assez précise de la répartition de la série. Par exemple on sait que pour une série quelconque la proportion des valeurs de la série en dehors de l'intervalle $[\bar{x} - 2s_x; \bar{x} + 2s_x]$ est inférieur à 25% et la proportion des valeurs de la série en dehors de l'intervalle $[\bar{x} - 3s_x; \bar{x} + 3s_x]$ est inférieur à 12%.