Chapitre 17 - Résumé statistiques

La fait de comptabiliser des bêtes, des populations, le cours des céréales, ... remonte au moins à 2300 ans avant JC en Chine. L'idée d'utiliser ces données pour prédire "le futur", construire des modèles est arrivée plus tardivement ; à la fin du XVIIIème siècle, on élabore des tables de mortalité pour estimer l'évolution de la population. C'est l'économiste allemand Gottfried Achenwall qui utilisa, vers 1749, le mot "STATISTIK" : la statistique représentant pour lui l'ensemble des connaissances que doit posséder un homme d'État.

L'idée de ce chapitre est de comparer efficacement des séries de données où les valeurs étudiées sont numériques en calculant des "paramètres".

I) Moyenne et écart-type

Moyenne : La moyenne, notée $\bar{x}$, s'obtient en additionnant les valeurs et en divisant par le nombre de valeurs.

Exemple : 1m57; 1m78; 1m60 ont pour moyenne $\frac{1,57+1,78+1,60}{3}=1,65$

Moyenne pondérée : C'est la moyenne que l'on peut calculer lorsqu'il y a des effectifs ou des coefficients.

Exemple : 3,5 et 0,5 coeff 4

âge	10	11
effectif	15	7

Moyenne d'âge des 22 élèves

$$moy=\frac{10\times15+11\times7}{15+7}\approx10,3$$

Propriété :

Si toutes les valeurs d'une série sont multipliées par un réel a alors la moyenne est aussi multipliée par a.

Si on augmente toutes les valeurs d'une série d'un réel b alors la moyenne est aussi augmentée de b.

En notant X la série de valeurs, et E la fonction calculant la moyenne d'une série (donc $E(X)=\bar{x}$), on a :

$$\forall a,b\in\mathbb{R},E(aX+b)=aE(X)+b$$

La moyenne ne suffit pas à représenter une série de donnée, on veut également considérer la dispersion des valeurs autour de la moyenne.

Par exemple, 1300 et 1500 ont une moyenne de 1400, mais 700 et 2100 aussi ont une moyenne de 1400. Pourtant il y a une différence notable entre ces deux mini-séries.

On va donc calculer l'écart positif de chaque valeur à la moyenne.
Pour calculer l'écart entre deux nombres, on les soustrait 1300-1400=-100 mais on prend la distance, donc la valeur positive, et $|1300-1400|=100$. Le problème, c'est que faire des calculs avec des valeurs absolues est compliqué (on ne peut pas enlever les valeurs absolues comme on veut), mais il y a un autre moyen d'enlever un signe -, en élevant au carré.

Variance :

La variance se calcule par : $\frac{n_1(x_1-\bar{x})^2+\ldots+n_k(x_k-\bar{x})^2}{n_1+\ldots+n_k}$

Elle mesure la dispersion des valeurs autour de la moyenne.

Mais comme on a élevé au carré, on aimerait bien l'enlever. Par exemple si les données sont des longueurs en cm, alors $(x_1-\bar{x})^2$ est en cm² , donc la variance n'est pas dans la même unité que les valeurs de notre série. Pour enlever le carré, on utilise ..... une racine carrée.

Ecart type :

C'est la racine carrée de la variance notée : $\sigma=\sqrt{V}$

Le couple $(\bar{x},\sigma)$ permet de résumer une série de donnée. Il a l'avantage de prendre en compte chacune des valeurs de la série. L'inconvénient est qu'il est très sensible aux valeurs extrêmes.
Il est d'usage de considérer les intervalles $[\bar{x}-\sigma,\bar{x}+\sigma]$ (qui contient 68% des valeurs), $[\bar{x}-2\sigma,\bar{x}+2\sigma]$ (qui contient 95% des valeurs) et $[\bar{x}-3\sigma,\bar{x}+3\sigma]$ (qui contient 99,7% des valeurs),

Voici un exemple : en orthophonie, on calcule, sur les résultats de tests similaires, l'écart-type correspondant. Suivant la valeur de l'écart-type (et suivant le type de test, présentant ou non une forte dispersions des résultats), on peut considérer :

	Zone pathologique	Moyenne basse (zone de fragilité)	Moyenne (zone dans la norme)	Moyenne haute (zone de confort)	Performance supérieure à la tranche d'âge
Déviations standard (D.S)	≤ -2	entre -1 et -2	entre -1 et +1	entre +1 et +2	> +2

Ainsi, on peut considéré qu'un trouble est pathologique lorsque les résultats sont à plus de deux écart-types de la moyenne.

II) Médiane et quartiles

Médiane : Lorsque les caractères sont classés dans l'ordre croissant, la médiane est une valeur qui partage en deux la série (l'effectif est le même avant et après la médiane)

Pour trouver la médiane, il faut diviser l'effectif total par 2 :

- si le résultat a une partie décimale non nulle, on arrondit au supérieur

exemple : $\frac{61}{2}=31,5$ donc la médiane est la 32ème valeur

- si le résultat est un entier n, la médiane se situe entre la nème valeur et la (n+1)ème.

exemple : $\frac{44}{2}=22$ donc la médiane est comprise entre la 22ème et la 23ème valeur.

Effectifs cumulés croissants :

Si l'effectif est trop grand, il est hors de question d'écrire toutes les valeurs pour trouver la médiane. On calcule alors les effectifs cumulés.

Nombre de frères et sœurs	0	1	2	3	4	5
Effectifs	2	3	10	8	4	2
Effectifs cumulés croissants	2	5	15	23	27	29

La ligne des ECC qu'on obtient en additionnant au fur et à mesure les effectifs:

2 2 + 3 = 5 2 + 3 + 10 = 15 etc...

Ici, l'effectif de la population est de 29. La moitié de 29 est 14,5. La médiane est la 15ème valeur, la médiane est donc 2.

Quartiles : On peut être plus précis et au lieu de couper en deux, on peut couper en 4, on obtient alors les quartiles.

Pour trouver Q_1, on divise l'effectif total par 4 *(ou on multiplie par 0,25)* et on arrondit au dessus.

Pour trouver Q_3, on multiplie l'effectif total par $\frac{3}{4}$ *(ou on multiplie par 0,75)* et on arrondit au dessus.

Nombre de connexions quotidiennes	1	2	3	4	5	6	7	8	9	10
Effectif	34	49	65	71	48	28	26	40	19	20
Effectifs cumulés croissants	34	83	148	219	267	295	321	361	380	400
		+49	+65	+71	+48	+28	+26	+40	+19	+20

L'effectif total est 400 :

$\frac{400}{4}=100$; le premier quartile est la centième valeur, c'est donc 3.

$\frac{2}{4}\times 400=200$; le second quartile (la médiane) est entre la $200^{ème}$ et la $201^{ème}$ valeur : c'est 4.

$\frac{3}{4}\times 400=300$; le troisième quartile est la $300^{ème}$ valeur ; c'est donc 7.

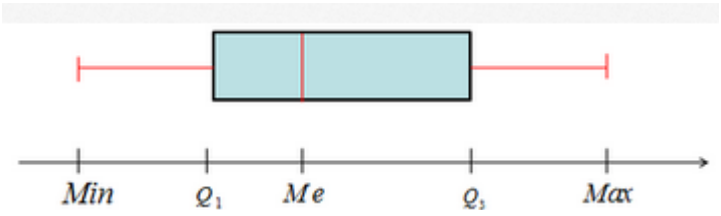
Ecart interquartile :

C'est le résultat de Q_3 - Q_1. Il permet de mesurer la dispersion des valeurs autour de la médiane.

Plus il est élevé, plus les valeurs sont dispersées.

Diagramme en boîte :

Aussi appelé diagramme de Tuckey, on schématise la série statistique étudiée par un dessin, composé de deux rectangles et deux "moustaches", sur un axe gradué.



Les diagrammes en boîtes permettent une comparaison visuelle aisée des séries statistiques.

Le couple médiane - quartiles permet de résumer une série de données efficacement. Il n'est pas sensible aux valeurs extrêmes et on peut le représenter avec un schéma.

Remarque :

Face à une série de données, quel couple choisir ?

En général on calcule les deux qui, si les données sont assez uniformes, sont très proches. Voici quelques disctinctions :

- Si vous avez des valeurs trop extrêmes, on préférera la médiane.
- La moyenne est influncée par toutes les valeurs de la série, pas la médiane qui n'est influencée que par le nombre de valeurs.
- Dans le cas général, les "gens" préfère la moyenne, c'est un concept plus facile à comprendre que la médiane dans le sens où la plupart considère que ce qui est dans la moyenne est "normal".
- Sachez enfin que la moyenne à le gros avantage de pouvoir se mettre à jour facilement (plus facilement que la médiane).

Imaginez que vous ayez eu 5 notes et que votre moyenne est, pour le moment, de 11. Si vous avez 16 comme 6ème note, quelle sera votre moyenne ?

Imaginez maintenant que vous ayez eu 5 notes et que votre médiane est, pour le moment, de 11. Si vous avez 16 comme 6ème note, quelle sera votre médiane ?

III) Algorithme

Avec les ordinateurs, on ne se pose plus trop la question et on calcule tout.

Voici deux programmes en python, vous devez être capable d'expliquer chaque ligne de ces programmes et leur but.

```
from math import sqrt
def param():
    m=(13+15+7+9+12+31+14+8+34+41+20+16+18+16+12+10+27+23+13+9)/20

    s= sqrt (((13-m)**2+(15-m)**2+(7-m)**2+(9-m)**2+(12-m)**2+(31-m)**2+(14-m)**2+
             (8-m)**2+(34-m)**2+(41-m)**2+(20-m)**2+(16-m)**2+(18-m)**2+(16-m)**2+
             (12-m)**2+(10-m)**2+(27-m)**2+(23-m)**2+(13-m)**2+(9-m)**2)/20)
    print(m,s)
```

```
from math import sqrt
L=[13,15,7,9,12,31,14,8,34,41,20,16,18,16,12,10,27,23,13,9]
def param(L):
    m=0
    s=0
    for a in L:
        m=m+a
    m=m/len(L)

    for a in L:
        s=s+(a-m)**2
    s=sqrt(s/20)

    p=0
    for a in L:
        if m-s<a<m+s:
            p=p+1
    p=p/20

    print(m,s,p)
```