Statistiques

I Présentation et Vocabulaire de base

Toute étude statistique s'appuie sur des données. Dans le cas ou ces données sont numériques (95% des cas), on distingue les données discrètes (qui prennent un nombre fini de valeurs : par ex, le nombre de voitures par famille en France) des données continues (qui prennent des valeurs quelconques : par ex, la taille des animaux d'un zoo).

- Dans le cas d'une série discrète, le nombre de fois ou l'on retrouve la même valeur s'appelle l'effectif de cette valeur. Si cet effectif est exprimé en pourcentage, on parle alors de fréquence de cette valeur.
- Dans le cas d'une série continue, on répartit souvent les données par classes.

Le but des statistiques est d'analyser les données dont on dispose. Pour cela, on peut par exemple chercher à déterminer la moyenne ou la médiane de la série. De tels nombres permettent notamment de comparer plusieurs séries entre elles. On les appelle indicateurs statistiques ou paramètres statistiques. On distingue les indicateurs de position (qui proposent une valeur "centrale" de la série) et les indicateurs de dispersion (qui indiquent si la série est très regroupée autour de son "centre" ou non).
Ainsi, le mode d'une série (valeur qui a le plus grand effectif de la série) est un indicateur de position.
L'étendue de cette série (différence entre la plus grande et la plus petite valeur) est un indicateur de dispersion.
La moyenne et la médiane sont des indicateurs de position.

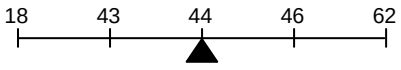
De plus, lorsque la série est trop importante (population d'un pays…), on est obligé de faire un sondage, c'est à dire de restreindre l'étude à un échantillon de cette série. Tout le problème est alors de choisir un échantillon vraiment représentatif (de taille suffisante et non biaisé) et d'évaluer l'erreur commise par rapport à une étude qui porterait sur l'ensemble de la série. (exemple des sondages électoraux)

II Médiane

Définition

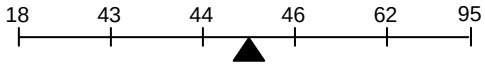
Soit une série statistique d'effectif total n, rangée par ordre croissant.
On appelle médiane la valeur "du milieu". On dit qu'elle partage la série en deux moitiés : il y a autant de valeurs en dessous qu'au dessus.
Pour déterminer son rang, il y a 2 cas :

- si n est impair : la médiane est la valeur de rang $\frac{n+1}{2}$



méd = 44

- si n est pair : nous prendrons la demi-somme des deux valeurs dont les rangs entourent le nombre $\frac{n+1}{2}$



méd = 45

Remarque :
Si les données ont été regroupées en classes, on ne peut déterminer la valeur exacte de la médiane. En revanche, on appellera classe médiane, la classe qui la contient (et permet donc d'en donner un encadrement).

Exemples
Données discrètes "en vrac"

10, 7, 12, 18, 16, 15, 5, 11, 11, 20, 15, 11, 18, 14

Ordonnons la série par ordre croissant : 5, 7, 10, 11, 11, 11, 12, 14, 15, 15, 16, 18, 18, 20

Il y a 14 termes or $\frac{14+1}{2} = 7{,}5$.

La médiane est donc la demi somme des $7^{ème}$ et $8^{ème}$ termes : méd $= \frac{12+14}{2} = 13$

Avec un tableau d'effectifs

valeurs	1	2	3	4	5	6
effectifs	6	11	25	19	15	5
effectifs cumulés	6	17	42	61	76	81

Attention, il faut bien interpréter cette dernière ligne : Les données qui valent 3 ont un rang compris entre 18 et 42 inclus

L'effectif total est de 81 or $\frac{81+1}{2} = 41$.

La médiane est donc le 41ème terme : méd = 3

Avec des données réparties par classes

classe	[0 ; 2[	[2 ; 4[	[4 ; 6[	[6 ; 8]
fréquence	10%	38%	45%	7%
fréquence cumulée	10%	48%	93%	100%

48% des valeurs sont strictement inférieures à 4
Et 93% des valeurs sont strictement inférieures à 6
La classe médiane est donc la classe [4 ; 6[
On peut donc en déduire l'encadrement suivant $4 < \text{méd} < 6$

III Moyenne

Exemple :

Soit la série statistique ci-contre :

valeurs	0	1	2	3	4
effectifs	1	2	1	4	2

La moyenne est : $\bar{x} = \frac{0+1+1+2+3+3+3+3+4+4}{1+2+1+4+2} = \frac{24}{10} = 2{,}4$

On préférera écrire : $\bar{x} = \frac{0+2\times1+2+4\times3+2\times4}{1+2+1+4+2} = \frac{24}{10} = 2{,}4$

Notation.

Le symbole Σ (sigma) signifie que l'on ajoute les éléments, par exemple

$$\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5$$

Formule pour le calcul de la moyenne

Soit la série statistique ci-contre :

valeurs	x_1	x_2	…	x_p
effectifs	n_1	n_2	…	n_p

La moyenne est : $\bar{x} = \frac{n_1x_1+n_2x_2+\ldots+n_{N-1}x_{N-1}+n_Nx_N}{n_1+n_2+\ldots+n_{N-1}+n_N} = \frac{\sum_{i=1}^{N} n_i x_i}{\sum_{i=1}^{N} n_i}$

Remarque :

Si les données ont été regroupées en classes, on ne peut calculer la valeur exacte de la moyenne. On peut toutefois en déterminer une bonne approximation en remplaçant chaque classe par son milieu.

Exemples :

a) Tableau de fréquences

valeurs	12	13	14	15	16
fréquences	0,05	0,17	0,43	0,30	0,05

$\overline{x}$ =

b) Données réparties en classes

classes	[0 ; 5[	[5 ; 10[	[10 ; 15[	[15 ; 20]
effectifs	7	12	14	2

Remplaçons chaque classe par son milieu :

$\overline{x} \approx$

IV Propriétés de la moyenne

a) Addition ou Multiplication de toutes les données par un même nombre :

Ex : Soit la série : 10, 12, 14. $\overline{x}$ =
Ajoutons 2 : la nouvelle série est : 12, 14, 16. $\overline{x}$ =
Multiplions par $\frac{1}{2}$: la nouvelle série est : 6, 7, 8. $\overline{x}$ =

Théorème de linéarité de la moyenne.
Si toutes les valeurs d'une série sont multipliées par un nombre *k* alors la moyenne est aussi multipliée par *k*.
Si on ajoute un nombre *k* à toutes les valeurs d'une série alors la moyenne est augmentée de *k*.

b) Moyennes partielles

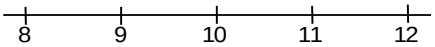
Ex : Sur les 5 premières interros, Paul a eu 12,5 de moyenne. Il vient d'avoir 15,5 à la 6ème interro.
Les notes ayant toutes le même coefficient, quelle est sa nouvelle moyenne ?
La somme des notes des 5 premières interros est : 12,5 × 5
La somme des notes des 6 interros est donc : 12,5 × 5 + 15,5
La nouvelle moyenne est donc : $\overline{x} = \frac{12{,}5 \times 5 + 15{,}5}{6} = 13$

Cas général : Si on réunis deux groupes disjoints ayant respectivement pour moyennes et effectifs, $\overline{x}_1$ et n_1 d'une part, $\overline{x}_2$ et n_2 d'autre part, la moyenne de l'ensemble sera alors :

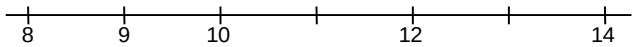
$$\overline{x} = \frac{n_1 \times \overline{x}_1 + n_2 \times \overline{x}_2}{n_1 + n_2}$$

c) Lien entre la moyenne et la médiane

- Quand on modifie les valeurs extrêmes d'une série, la moyenne change contrairement à la médiane qui ne change pas. On dit que la moyenne est "sensible aux valeurs extrêmes".
  Il arrive que certaines de ces valeurs extrêmes soient douteuses ou influent de façon exagérée sur la moyenne. On peut alors, soit calculer une moyenne élaguée (c'est à dire recalculer la moyenne sans ces valeurs gênantes), soit utiliser la médiane.
- Comment interpréter un écart entre la moyenne et la médiane ?

  Soit la série suivante : 

  Ici la moyenne (10) et la médiane (10) sont identiques : la série est bien "centrée".

  Soit la nouvelle série : 

  Ici la moyenne (10,6) est plus importante que la médiane (10) : la série est plus "étalée à droite".

V Quartiles

Définitions :

Le premier quartile Q_1est la plus petite des valeurs de la série telle qu'au moins 25% de la population ait sa valeur inférieur ou égale à Q_1.

Le troisième quartile Q_3est la plus petite des valeurs de la série telle qu'au moins 75% de la population ait sa valeur inférieure ou égale à Q_3.

Méthode pour trouver Q_1 et Q_3.

Pour une population d'effectif n, $\frac{n+1}{4}$ (ou $\frac{3(n+1)}{4}$) si il est un entiers nous donne le rang de Q_1 (ou Q_3), dans le cas contraire on prendra la valeur de l'élément du rang immédiatement supérieur.

Exemple

Avec un tableau d'effectifs

valeurs	1	2	3	4	5	6
effectifs	6	11	25	19	15	5
effectifs cumulés	6	17	42	61	76	81

Ici n = 81

Donc $\frac{n+1}{4} = 20{,}5$ il va donc falloir que je prenne la valeur de l'individu de rang 21, donc $Q_1 = 3$

$\frac{3(n+1)}{4} = 61{,}5$ L'individu de rang 62 a pour rang 5 donc $Q_3 = 5$

VI Mode et classe modale

Définition

Un mode est une valeur du caractère ayant le plus grand effectif.

Une classe modale est une classe du caractère ayant le plus grand effectif.

Dans l'exemple précédent :

Le mode est 3 car l'effectif correspondant est 25 ce qui est le plus gros effectif

Dans le III exemple b) la classe modale est [10 ;15[car son effectif , 14, est le plus grand des effectifs.

VII Fluctuation d'échantillonnage et simulation

Soit une population dont la proportion d'apparition d'une caractéristique donnée est p (pϵ[0,1]).

Lorsque l'on considère un échantillon de taille n de cette population, si il est représentatif, f sa fréquence d'apparition de la caractéristique aura 95% de chance d'être dans l'intervalle $[p - \frac{1}{\sqrt{n}}, p + \frac{1}{\sqrt{n}}]$.

Si un échantillon présente une fréquence n'étant pas dans l'intervalle, on peut raisonnablement l'éliminer.

Inversement

Lors d'un sondage, on ne peut interroger toute la population, on se concentrera sur un échantillon non biaisé, si la fréquence d'apparition du caractère est f alors on peut être sur à 95% que dans la population la proportion d'apparition de la caractéristique est dans « l'intervalle de confiance à 95% » : $[f - \frac{1}{\sqrt{n}}, f + \frac{1}{\sqrt{n}}]$

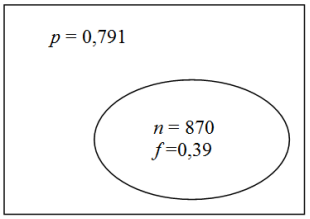
Ces intervalles peuvent nous aider à éliminer les échantillons biaisé lorsque l'on connaît p , et à estimer p si on sait que l'échantillon est non biaisé.

Exemple d'utilisation de l'intervalle de confiance.

En Novembre 1976 dans un comté du sud du Texas, Rodrigo Partida était condamné à huit ans de prison pour cambriolage d'une résidence et tentative de viol.

Il attaqua ce jugement au motif que la désignation des jurés de ce comté était discriminante à l'égard des Américains d'origine mexicaine. Alors que 79,1% de la population du comté était d'origine mexicaine, sur les 870 personnes convoquées pour être jurés lors des 11 années précédentes, seules 339 d'entre elles étaient d'origine mexicaine.

Est-ce que le jury est représentatif de la population ?



Recherchons l'intervalle de confiance $[\ p - \frac{1}{\sqrt{n}},\ p + \frac{1}{\sqrt{n}}]$
Ici il sera donc de $[0{,}791 - \frac{1}{\sqrt{870}}\ ;\ 0{,}791 + \frac{1}{\sqrt{870}}] \approx [0{,}76\ ;\ 0{,}82]$
Ici la fréquence est d'environ 0,39 elle n'est donc pas dans l'intervalle de confiance. Le jury n'est donc pas représentatif de la population du comté.

Exemple de synthèse
On s'intéresse aux deux classes de secondes qui ont M. Kergot comme prof de maths
Les notes de la seconde 4 au DS1 de mathématiques sont :
3,5 ;5 ;5 ;8 ;11 ;11,5 ;12 ;12,5 ;12,5 ;14 ;14 ;15 ;15,5 ;16 ;16,5 ;16,5 ;17 ;17 ;17,5 ;18,5 ;19 ;19 ;20
Les notes de la seconde 14 au DS1 de mathématiques sont :
3,5 ;10,5 ;11,5 ;11,5 ;12,5 ;12,5 ;13 ;14,5 ;14,5 ;14,5 ;15 ;15 ;15,5 ;16 ;17 ;17 ;
17,5 ;18 ;18 ;18,5 ;19 ;19 ;19 ;19 ;19 ;19,5 ;20 ;20 ;20 ;20 ;20 ;20 ;20
1) Déterminer pour chaque classe l'effectif, moyenne, médiane, Q1, Q3, étendue
2) Calculer la moyenne de tous les élèves secondes de M. Kergot
3) quelle est la fréquence du caractère : l'élève a 12 ou moins dans chacune des classes.
4) on prend un échantillon 4 élèves de seconde 4, trois quart d'entre eux ont eu 12 ou moins, l'échantillon est il représentatif ?
5) on prend un échantillon 16 élèves de seconde 14, la moitié d'entre eux ont eu 12 ou moins, l'échantillon est il représentatif ?

Réponses :
1) *seconde 4* — *23 élèves, moyenne* $\frac{316{,}5}{23} \approx 13{,}76086957$*;*
(23+1)/2 = 12 la médiane est de 15
(23+1)/4 = 6 la note de rang 6 est : 11,5 donc Q1 = 11,5
3(23+1)/4 = 18 la note de rang 18 est : 17 donc Q3 = 17
Etendue : 20 – 3,5 = 16,5

Seconde 14 — *33 élèves, moyenne* $\frac{540{,}5}{33} \approx 16{,}37878788$ *;*
(33+1)/2 = 17 la médiane est de 17,5
(33+1)/4 = 8,5 la note de rang 9 est : 14,5 donc Q1 = 14,5
3(33+1)/4 = 25,5 la note de rang 26 est : 14,5 donc Q3 = 19,5
Etendue : 20 – 3,5 = 16,5

2) $\bar{x} = \frac{\frac{540{,}5}{33}\times 33 + \frac{316{,}5}{23}\times 23}{56} = \frac{857}{56} \approx 15{,}30$
3) nombre d'élèves ayant eu 12 ou moins :
7 élèves sur 23 en seconde 4, $p_4 \approx 0{,}304$ 4 élèves sur 33 en seconde14 , $p_{14} \approx 0{,}121$
4) ici f = 0,75 Recherchons l'intervalle de confiance $[p_4 - \frac{1}{\sqrt{n}},\ p_4 + \frac{1}{\sqrt{n}}]$
Ici il sera donc de $\left[0{,}304 - \frac{1}{\sqrt{4}}; 0{,}304 + \frac{1}{\sqrt{4}}\right] \approx$ [-0,195 ; 0,805] donc 0,75 est bien dans l'intervalle de confiance à 95%, c'est donc un échantillon représentatif.
5) ici $f = 0{,}5$ Recherchons l'intervalle de confiance $[p_{14} - \frac{1}{\sqrt{n}},\ p_{14} + \frac{1}{\sqrt{n}}]$
Ici il sera donc de $\left[0{,}121 - \frac{1}{\sqrt{16}}; 0{,}121 + \frac{1}{\sqrt{16}}\right] \approx$ [-0,129 ; 0,371] donc n'est pas dans l'intervalle, donc on peut dire que ce dernier n'est pas représentatif de la classe.

Conclusion :
Si on prend un échantillon très petit, des résultats aberrants peuvent être validés, donc il faudra prendre des échantillons de taille raisonnable.