V) Adéquation de données à une loi équirépartie

1) Statistiques

Vocabulaire

La statistique est une discipline scientifique qui étudie quantitativement les CARACTERES des INDIVIDUS d'une POPULATION.

Le CARACTERE peut – être QUALITATIF ou QUANTITATIF.

Les VALEURS du CARACTERE sont appelées LES VARIABLES ou LES MODALITES.

Définition

Un caractère est QUANTITATIF lorsque les variables sont numériques

Un caractère est QUALITATIF lorsque les variables NE sont PAS numériques

Exemples :

Dans une entreprise de construction de bungalows on veut étudier les commandes des clients, et plus particulièrement la structure et la surface des bungalows commandés.

Population : Ensemble des bungalows livrés.
Individu : Un bungalow
Caractère quantitatif : Structure des bungalows
Variables : Béton (parpaings) , béton (voile), béton et bois, bois .
Ici on a obtenu une SERIE QUALITATIVE.
Caractère qualitatif : surface des bungalows
Variables : 20 m^2 ; 25m^2 ; 30 m^2 ; 35 m^2.
Ici on a obtenu une série QUANTITATIVE DISCRETE.

Etude d'une série statistique quantitative discrète
Exemple :
On étudie les notes de mathématiques de deux classes de TS obtenues à un examen blanc.
Dans la classe de TS 1 on a les résultats suivants :
2 , 3, 7, 7, 10, 11, 14, 16, 16, 17

Dans la classe de TS 2 on a les résultats suivants :
8, 8, 8, 9, 10, 10, 11, 11,11, 11, 16

On voit que dans la classe de TS 1 les élèves obtiennent 2 fois la note 7, l'effectif de la note 7 est donc 2. Il ya 10 élèves donc **l'effectif total** est 10,

le quotient $\frac{2}{10}$ est **la fréquence** de la note 7.

Pour chaque série quantitative on va résumer les résultats dans un tableau :

Classe de TS 1

Variables x_i	2	3	7	10	11	14	16	17	Total
Effectifs n_i	1	1	2	1	1	1	2	1	10
Fréquences f_i	10%	10%	20%	10%	10%	10%	20%	10%	1

Exercice : compléter le tableau

Classe de TS 2

Variables x_i						Total
Effectifs n_i						
Fréquences f_i						

CAS GENERAL

Soit la série statistique S quantitative résumée dans le tableau suivant :

Variables x_i	x_1	x_2	…	x_p	TOTAL
Effectifs n_i	n_1	n_2	…	n_p	N
Fréquences f_i	$f_{1=}\frac{n_1}{N}$	$f_{2=}\frac{n_2}{N}$	…	$f_{p=}\frac{n_P}{N}$	1

Remarque : $N = n_1 + n_2 + \ldots + n_p$

Lorsque les valeurs sont nombreuses, le tableau peut – être interprété à l'aide de quelques valeurs caractéristiques numériques. La plus simple et la plus connue est la moyenne.

Définition

La moyenne notée $\bar{x}$ de la série statistique S est :

Formule avec les effectifs : $\bar{x} = \dfrac{n_1 x_1 + n_2 x_2 + \ldots + n_p x_p}{N}$

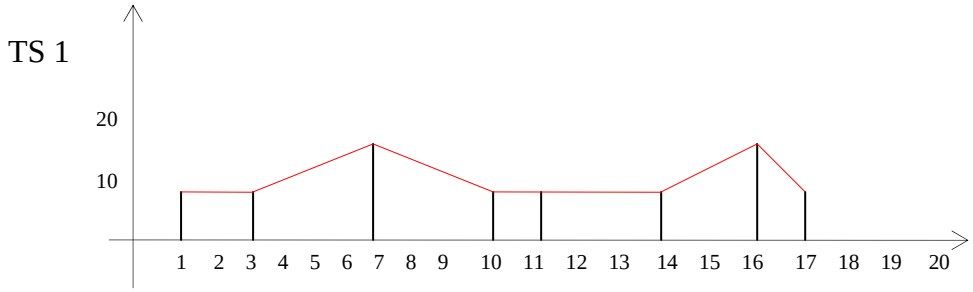
Formule avec les fréquences : $\bar{x} = f_1 x_1 + f_2 x_2 + \ldots + f_p x_p$

Exemple : dans les deux classes de TS 1 la moyenne de mathématiques est $\bar{x}_1 = 10,3$.

et de TS 2 $\bar{x}_2 \approx 10,27$.

Représentation graphique : diagrammes en bâtons

Chaque donnée est représentée par un bâton dont **la hauteur est proportionnelle à l'effectif ou à la fréquence**. On peut compléter le diagramme en dessinant le polygone des effectifs . Il est formé des segments joignant les sommets des bâtons.

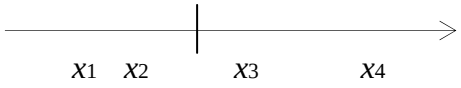


MEDIANE :

Dire que M_e est la médiane d'une série statistique signifie qu'il y a autant de valeurs de la série supérieures à M_e que de valeurs de la série inférieures à M_e.

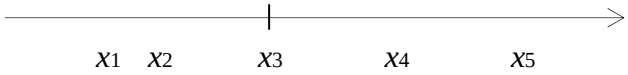
Si la série statistique a pour effectif total N :

- Avec N PAIR : $M_e = (x_2+x_3)/2$



La médiane est demi – somme des deux valeurs centrales, $[x_2 ; x_3]$ est l'intervalle médian.

- Avec N IMPAIR : $M_e = x_3$



La médiane est la valeur centrale.

Dans le cas de la série du TS1 : 2 , 3, 7, 7, 10, 11, 14, 16, 16, 17

.....

Dans le cas de la série du TS 2 : 8, 8, 8, 9, 10, 10, 11 , 11,11, 11, 16

.....

QUARTILES

La médiane M_e sépare une série statistique en deux sous- séries de même effectif, l'une contenant les plus petites valeurs, l'autre les plus grandes.

LES QUARTILES sont les médianes de ces sous – séries.

Le premier quartile noté Q_1 est la médiane de la sous - série inférieure, on dit aussi que au moins un quart (ou 25%) des valeurs prises sont inférieures ou égales à Q_1.

Le troisième quartile noté Q_3 est la médiane de la sous – série supérieure, on dit aussi que au moins trois quarts (ou 75%) des valeurs prises sont supérieures ou égales à Q_3.
A Noter : La médiane M_e est parfois appelée aussi deuxième quartile.

Définition

Le nombre $Q_3 - Q_1$ est appelé ECART INTERQUARTILE.
L'intervalle $]Q_1 ; Q_3[$ est appelé INTERVALLE INTERQUARTILE.

Remarques :

1. Les nombres Q_1 , M_e , Q_3 permettent de couper la population étudiée en quatre groupes contenant chacun le même nombre d'éléments.
2. **NOTION DE DECILES** : on peut définir de façon analogue la notion de déciles qui sont des nombres qui permettent de couper la population en dix groupes contenant chacun le même nombre d'éléments.

On utilise alors **D_1 le premier décile, et D_9 le neuvième décile**.
Au moins un dixième (ou 10%) des valeurs prises sont inférieures ou égales à D_1
Au moins neuf dixièmes (ou 90%) des valeurs prises sont supérieures ou égales à D_9.

Exemples :
Dans le cas de la série du BTS1 : 2 , 3, 7, 7, 10, 11, 14, 16, 16, 17

$Q_1 = 7$ $Q_3 = 16$

Dans le cas de la série du BTS 2 : 8, 8, 8, 9, 10, 10, 11 , 11,11, 11, 16

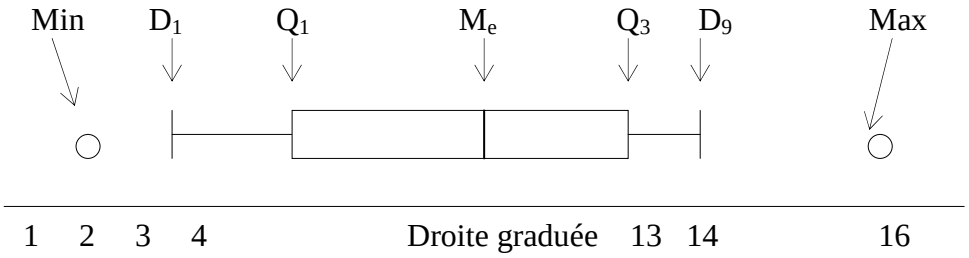
Exemples pour les déciles :
On considère deux entreprises A et B de 100 salariés dont la grille des salaires est décrite par les séries statistiques ci – dessous :

Salaire mensuel En milliers d'euros	1	1.1	1.2	1.3	1.4	1.5	2	2.2	2.4	2.5
Effectifs Entreprise A	25	20	15	15	8	7	6	2	1	1
Effectifs Entreprise B	20	15	12	10	10	7	8	8	6	4

Pour A : $D_1 = Q_1 = 1$ $Q_3 = 1.3$ $D_9 = 1.5$ $M_e = 1.2$
Pour B : $D_1 = 1$ $Q_1 = 1.1$ $Q_3 = 2$ $D_9 = 2.2$ $M_e = 1.3$

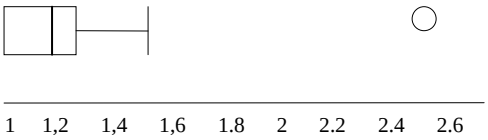
DIAGRAMMES EN BOITE

Pour utiliser les quartiles et la médiane on utilise une représentation simple appelée « diagramme en boîte » ou « diagramme à moustaches ».

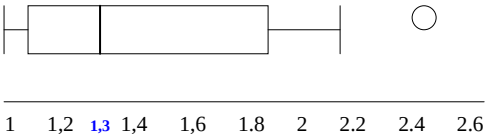


Dans nos exemples des entreprises A et B :

Entreprise A



Entreprise B



Exercices : Construire les diagrammes en boîtes des séries des TS 1 et 2

2°) Adéquation : Etude d'un exemple et propriété

Un joueur veut vérifier si le dé qu'il possède est « normal », c'est – à – dire bien équilibré.

Pour cela on lance un dé cubique 200 fois et on note les résultats obtenus :

Variables x_i	1	2	3	4	5	6	Total
Effectifs n_i	31	38	40	32	28	31	200
Fréquences f_i	0.155	0 ,19	0,2	0,16	0,14	0,155	1

Pour savoir si la distribution de fréquences obtenues est « proche » de la loi uniforme,
(car on sait que quand le dé est équilibré $P(\{1\}) = P(\{2\}) = \ldots = P(\{6\}) = 1/6$)

On calcule la quantité suivante, qui prend en compte l'écart existant entre chaque fréquence trouvée et la probabilité théorique attendue :

$$d_0^{\,2} = (0{,}155 - 1/6)^2 + (0{,}19 - 1/6)^2 + (0{,}2 - 1/6)^2 + (0{,}16 - 1/6)^2 + (0{,}14 - 1/6)^2 + (0{,}155 - 1/6)^2$$

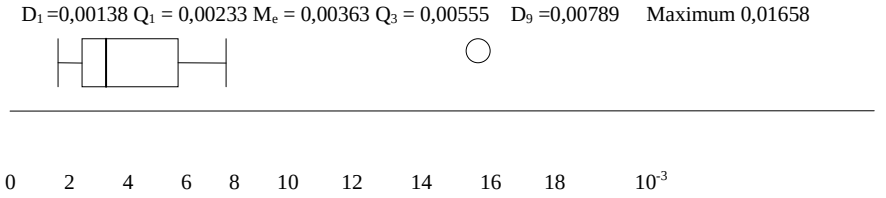
$$d_0^{\,2} \approx 0.00268$$

Cependant cette valeur en elle – même, bien qu'elle semble petite ne nous renseigne pas.
En effet ici les résultats obtenus dépendent des lancers effectués, si l'on effectue 200 autres lancers on aura d'autres résultats pour les fréquences : C'est ce que l'on appelle la **fluctuation d'échantillonnage**.

Pour décider si d^2 est suffisamment proche de la valeur qui permettrait d'accepter ou de rejeter

l'hypothèse d'équiprobabilité, on simule un grand nombre de fois l'expérience de 200 lancers d'un dé qui serait parfaitement équilibré (à l'aide d'un programme aléatoire sur tableur, calculatrice…) et on étudie la série des valeurs obtenues pour d^2.

Les résultats pour 1000 simulations de 200 lancers d'un dé qui serait équilibré sont résumés A l'aide du diagramme à moustache ci - dessous :



Le neuvième décile de la série simulée de d^2 est 0,00789.
Cela signifie que 90% des valeurs de d^2 obtenues au cours de ces 1000 simulations sont dans I = [0 ; 0,00789]. Or $d_0^2 < 0,00789$ soit d_0^2 dans I, ce qui veut dire que le dé est équilibré mais que l'on a tout de même 10% de chances de se tromper. On dit que le
Seuil de confiance est de 90% ou que le seuil de risque est de 10%.

Propriété

Soit une épreuve conduisant aux issues $a_1, a_2, \dots, a_q$.
Expérimentalement, si on répète n fois cette épreuve ($n \geq 100$) , on obtient les fréquences $f_1, f_2, \dots, f_P$ pour chacune des issues. Pour vérifier l'adéquation de ces données à la loi équirépartie sur $\{ a_1, a_2, \dots, a_q\}$, on calcule le nombre $d^2 = (f_1 - 1/q)^2 + \dots + (f_q - 1/q)^2$.

La réalisation d'un grand nombre de simulations de cette épreuve conduit pour la variable d^2 à une série statistique de neuvième décile D_9 .

Si $d^2 \leq D_9$, alors on dira que les données sont compatibles avec le modèle de la loi uniforme Au seuil de risque de 10%.
Si $d^2 > D_9$, on dira que les données ne sont pas compatibles avec ce modèle au seuil de risque de 10%.

Exemple :

Le tireur décide de tester le dé tétraédrique afin de savoir s'il est bien équilibré ou s'il est pipé. Pour cela il lance 200 fois ce dé et il obtient le tableau suivant :

Face k	1	2	3	4
Nombre de sorties de la face k	58	49	52	41

a. Calculer les fréquences de sorties f_k observées pour chacune des faces.

b. On pose $d^2 = \Sigma_{k=1}^{4} \left(f_k - \frac{1}{4}\right)^2$. Calculer d^2.

c. On effectue maintenant 1000 simulations des 200 lancers d'un dé tétraédrique bien équilibré et on calcule pour chaque simulation le nombre d^2. On obtient pour la série statistique des 1000 valeurs de d^2 les résultats suivants :

Minimum	D_1	Q_1	Médiane	Q_3	D_9	Maximum
0,00124	0,00192	0,00235	0,00281	0,00345	0,00452	0,01015

Au risque de 10 %, peut-on considérer que ce dé est pipé ?

BAC JUIN 2006