Séance du 27.11. - Présentation - Statistiques

Ben Hassine Youssef	Halidi Zaina
Bucher Lucas	Martin Valentin
Bytyqi Genis	Neron De Azevedo Aloys
Coulon Ethan	Perananthasivan Shanthos
Deiber Edouard	Sapmaz Ceyhun
Deike Mathieu	Schumacher Mathias
Guzel Emre	Soibaha Fahim
	Sahanoglu Sinem

1. Statistiques à une variable
2. Statistiques à deux variables
3. Généralités sur les fonctions
4. Dérivation
5. Probabilités élémentaires
6. Calcul intégral, limites
7. Variables aléatoires discrètes
8. Variables aléatoires continues
9. Fiabilité, loi exponentielle
10. Opérations sur les variables aléatoires
11. Suites numériques

STATISTIQUES A UNE VARIABLE

I. Définitions et notations

I.1. Notations

Intro

Dans ce chapitre, on étudie des séries à caractères **quantitatifs** discrètes (à valeurs séparées) ou **continues** (dont les valeurs sont regroupées en *classes*, ou *intervalles*).
Dans le cas d'une série continue, on fait toujours l'hypothèse que la *répartition* des valeurs est *uniforme* à l'intérieur de chaque classe.

Notations

- X est le caractère étudié et $x_1, x_2, \ldots, x_p$ les *valeurs* du caractère ou les centres des classes dans le cas où les valeurs sont regroupées en classes. Ces valeurs sont rangées dans l'ordre croissant.
- $n_1, n_2, \ldots, n_p$ sont les *effectifs* respectifs des valeurs $x_1, x_2, \ldots, x_p$.
- $f_1, f_2, \ldots, f_p$ sont les *fréquences* respectives des valeurs $x_1, x_2, \ldots, x_p$.
- N est l'effectif total, et $N = n_1 + n_2 + \cdots + n_p$

X : caractère étudié : Distance domicile - lycée

Classes	x_i	n_i	f_i	Effectifs cumulés croissants	Effectifs cumulés décroissants	Fréquences cumulées croissantes
$[0;5[$	2,5	5	$=\frac{5}{12}\approx 0{,}42$	5	12	0,42
$[5;10[$	7,5	1	0,08	6	7	0,5
$[25;30[$	27,5	2	0,17	8	6	0,67
$[40;45[$	42,5	3	0,25	11	4	0,92
$[55;60[$	57,5	1	0,08	12	1	1
	Total	12	1			

I.2. Fréquences

Propriétés

La fréquence de la valeur x_i est donnée par $f_i = \frac{n_i}{N}$.
Ce nombre est compris dans l'intervalle $[0;1]$ et est souvent écrit sous la forme d'un *pourcentage*.
La somme de toutes les fréquences est égale à 1 : $\sum_{i=1}^{p} f_i = f_1 + f_2 + \cdots + f_p = 1$.

I.3. Effectif cumulé, fréquence cumulée

Définitions - Cas d'une série discrète

C'est le nombre d'individus pour lesquels la valeur du caractère est *inférieure ou égale* à x_i.
On a $S_i = n_1 + n_2 + \cdots + n_i$ et de ce fait on a $S_p = N$ (dernier effectif cumulé).
La fréquence cumulée en x_i est donnée par $F_i = f_1 + f_2 + \cdots + f_i = \frac{S_i}{N}$.

Définition - Cas d'une série continue

Pour l'effectif cumulé d'une classe $]a;b]$, il s'agit de l'effectif cumulé en b (borne droite de l'intervalle), c'est-à-dire le nombre d'individus pour lesquels la valeur du caractère est inférieure ou égale à b.
Ainsi, lorsqu'on calcule des effectifs cumulés pour une série statistique groupée en classes, ces effectifs cumulés doivent être impérativement affectés aux *bornes de droite* de ces classes.

Propriété - Hypothèse de répartition uniforme

La distribution de l'effectif de la classe $]a;b]$ est *uniforme* signifie en pratique que, si c et d sont deux nombres tels que $a \leqslant c < d \leqslant b$, alors l'effectif de l'intervalle $]c;d]$ vérifie la relation de proportionnalité :

$$\frac{\text{effectif de }]c;d]}{d-c} = \frac{\text{effectif de }]a;b]}{b-a}.$$

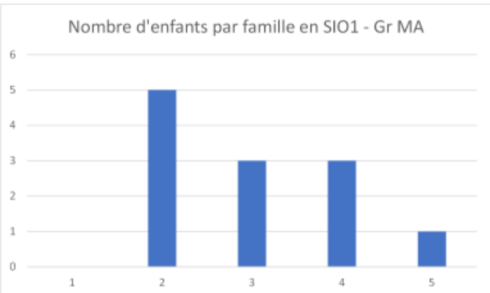
I.4. Représentations graphiques classiques

Définition - Diagramme en bâtons

Dans le cas d'une série discrète, on peut faire figurer les valeurs x_i du caractère sur l'axe des abscisses, et de tracer, à partir de chaque point de coordonnées$(x_i;0)$, un segment vertical dont la hauteur est proportionnelle à l'effectif n_i (ou à la fréquence) de la valeur x_i.

Y : Etude du nombre d'enfants par famille (variable discrète)

x_i	n_i	f_i
1	0	0
2	5	0,42
3	3	0,25
4	3	0,25
5	1	0,08
Total	12	1



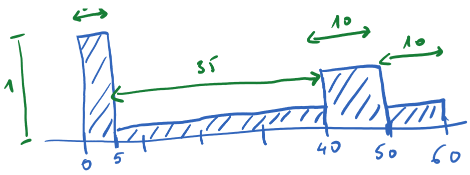
Définition - Histogramme

Dans le cas d'une série continue, les effectifs sont représentés par des **surfaces**.
Pratiquement, on construit sur chaque classe $]a_i;a_{i+1}]$ de l'axe des abscisses un *rectangle* dont l'aire est proportionnelle à l'effectif n_i de la classe concernée.

Classes	x_i	n_i	f_i
[0; 5[	2,5	5	$= \frac{5}{12} \approx 0{,}42$
[5; 40[	22,5	3	0,25
[40; 50[	45	3	0,25
[50; 60[	55	1	0,08

Aire rect.1 : $5 = l_1 \times h_1 = 5 \times h$ donc $h_1 = 1$
Aire rect.2 : $3 = l_2 \times h_2 = 35 \times h_2$ donc $h_2 = \frac{3}{35}$

[5; 40[	22,5	3	0,25
[40; 50[	45	3	0,25
[50; 60[	55	1	0,08
	Total	12	1



II. Paramètres d'une série

II.1. Médiane et quartiles

Définition

La valeur médiane d'une série statistique est la valeur M séparant la population en deux moitiés : les 50 % ayant une valeur inférieure ou égale à M et les 50 % ayant une valeur supérieure ou égale à M.

Exemple :

3
5
8
8
9
10
11
12
12
13
14
15
15
18
20

7	
8	
9	
11	
11	
11	
11	11,5
12	
12	
13	
14	
15	
18	
20	

1ère série : $N = 15$

Q_1 : $N \times \frac{1}{4} = 3,75$

Q_1 est la 4e valeur

$Q_1 = 8$.

Q_3 : $N \times \frac{3}{4} = 11,25$

Q_3 est la 12e valeur

$Q_3 = 15$

Méthodes

Cas d'une série **discrète** :
Si l'effectif est impair, la médiane est la valeur de rang $(N+1)/2$; si l'effectif est pair, la médiane est la moyenne des valeurs de rang $N/2$ et $N/2+1$.

Cas d'une série **continue** :
La médiane est la valeur correspondant à la fréquence cumulée égale à 0,5.

Définitions

De même, le premier quartile est la valeur Q_1 telle que 25 % de la population a une valeur du caractère inférieure ou égale à Q_1, les 75 % restants ayant une valeur supérieure ou égale à Q_1.
Le troisième quartile est la valeur Q_3 telle que 75 % de la population a une valeur du caractère inférieure ou égale à Q_3, les 25 % restants ayant une valeur supérieure ou égale à Q_3.

Méthodes

Cas d'une série **discrète** :
Si $N/4$ est un entier, Q_1 est la valeur de rang $N/4$ et Q_3 est la valeur de rang $3N/4$.
Si $N/4$ n'est pas un entier, Q_1 est est la valeur de rang immédiatement supérieur à $N/4$ et Q_3 est la valeur de rang immédiatement supérieur à $3N/4$.

Cas d'une série **continue** :
Q_1 est la valeur correspondant à la fréquence cumulée égal à 0,25 et Q_3 est la valeur correspondant à la fréquence cumulée égale à 0,75.

Définitions

Intervalle interquartile : c'est l'intervalle $[Q_1 ; Q_3]$. Écart interquartile : c'est le nombre $Q_3 - Q_1$.

II.2. Moyenne

Définition

Pour une série **discrète**, la moyenne des valeurs $x_1, x_2, \ldots, x_p$ d'effectifs $n_1, n_2, \ldots, n_p$ (eff. total N) est le réel :

$$\bar{x} = \frac{n_1x_1 + n_2x_2 + \cdots + n_px_p}{n_1 + n_2 + \cdots + n_p} = \frac{1}{\mathrm{N}}\sum_{i=1}^{p} n_i x_i.$$

Pour une série **continue**, on applique la même formule en prenant pour les valeurs x_i les centres des classes.

II.3. Variance et écart-type

Intro

La moyenne d'un caractère statistique ne suffit pas pour caractériser le comportement de ses valeurs.
Par exemple, les étudiants d'un premier groupe peuvent avoir obtenu des résultats très homogènes (c'est-à-dire des notes comprises dans un petit intervalle autour de 10) alors que ceux d'un deuxième groupe peuvent avoir des résultats beaucoup plus dispersés autour de 10. L'**écart-type** est un nombre qui mesure de cette **dispersion**, au sens où plus l'écart-type est élevé plus les valeurs sont dispersées autour de la moyenne.

Définitions

La **variance** d'une série statistique de caractère X prenant les valeurs (ou centres des classes) $x_1, x_2, \ldots, x_p$ d'effectifs respectifs $n_1, n_2, \ldots, n_p$ et de moyenne $\bar{x}$ est le nombre

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

On en déduit l'**écart-type** de la série $\sigma(\mathrm{X}) = \sqrt{\mathrm{V}(\mathrm{X})}$ qui est toujours un nombre positif.

Exemples "à la main" :

x_i	n_i	f_i	Fréquences cumulées croissantes	$(x_i - m)^2$	n_i
11	2	0,25	0,25	1,5625	2
11	1	0,125	0,375	1,5625	1
12	3	0,375	0,75	0,0625	3
14	1	0,125	0,875	3,0625	1
15	1	0,125	1	7,5625	1
Total	8			Total	8

Moyenne	12,25
Variance	1,94
Ecart-type	1,39

x_i	n_i	f_i	Fréquences cumulées croissantes	$(x_i - m)^2$	n_i
5	2	0,25	0,25	52,5625	2
6	1	0,125	0,375	39,0625	1
15	3	0,375	0,75	7,5625	3
18	1	0,125	0,875	33,0625	1
19	1	0,125	1	45,5625	1
Total	8			Total	8

Moyenne	12,25
Variance	30,69
Ecart-type	5,54

Point histoire

L'écart type est une grandeur dont l'invention remonte au XIXe siècle, qui vit la statistique se développer au Royaume-Uni. C'est à *Abraham de Moivre* (1667/1754,) qu'est attribuée la découverte du concept de mesure de la dispersion qui apparaît dans son ouvrage *The Doctrine of Chances* en 1718.

Mais le terme d'écart type (« standard deviation ») a été employé pour la première fois par *Karl Pearson* (1857/1936,) en 1893 devant la Royal Society. C'est aussi lui qui utilise pour la première fois le symbole σ pour représenter l'écart type.

III. Représentations graphiques

Intro

Il existe, comme présenté en cours, plusieurs représentations graphiques possibles pour des séries statistiques.
Pour des séries à valeurs discrètes, on peut utiliser :

- le nuage de points ou la « courbe » ;
- le diagramme en bâtons ou en barres ;
- le diagramme circulaire.

Pour des séries à valeurs continues, on va utiliser :

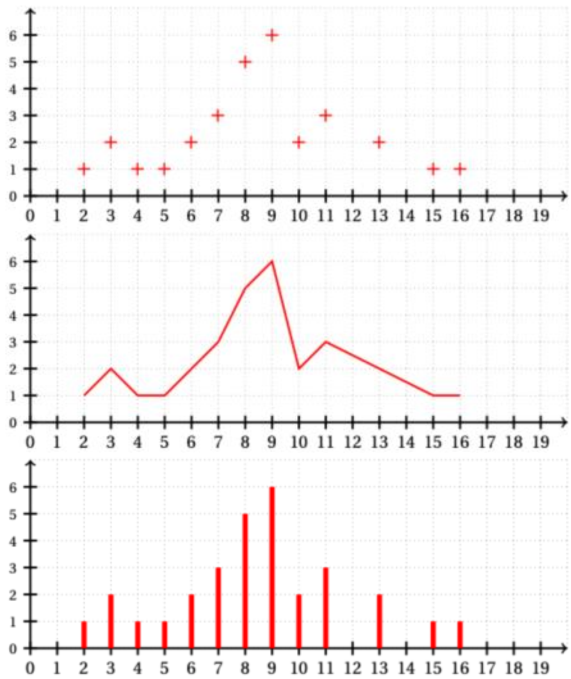
- le diagramme circulaire ;
- l'histogramme.

Exemple

On considère la série (discrète) suivante :

Notes	02	03	04	05	06	07	08	09	10	11	13	15	16
Eff.	1	2	1	1	2	3	5	6	2	3	2	1	1

Illustrations



Exemple

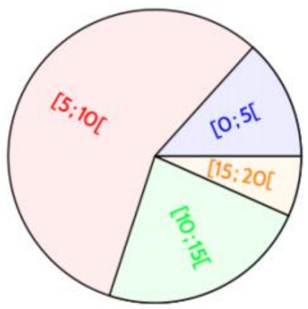
On considère la série (continue) suivante :

Notes	[0 ; 5[	[5 ; 10[	[10 ; 15[	[15 ; 20[
Effectif	4	17	7	2

Illustration

Pour le digramme circulaire, on travaille par proportionnalité par rapport à « 360° $\rightleftarrows$ 30 (effectif total) ».

- effectif de 4 : angle de $\frac{360 \times 4}{30} = 12 \times 4 = 48°$;
- effectif de 17 : angle de $\frac{360 \times 17}{30} = 12 \times 17 = 204°$;
- effectif de 7 : angle de $\frac{360 \times 7}{30} = 12 \times 7 = 84°$;
- effectif de 2 : angle de $\frac{360 \times 2}{30} = 12 \times 2 = 24°$.

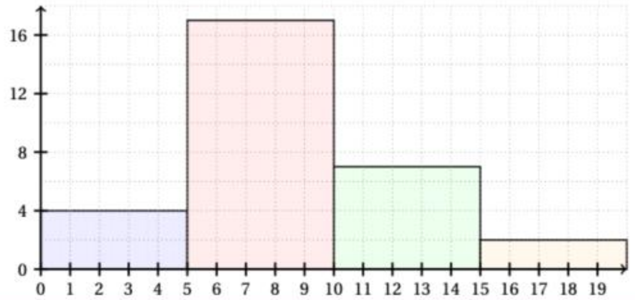


Rappel

Pour l'histogramme, c'est l'**aire** de chaque rectangle qui est proportionnelle à l'effectif (ou à la fréquence) associée à chaque classe.
Lorsque les classes ont la même **amplitude**, c'est la *hauteur* qui est proportionnelle à l'effectif.

Illustration

Ici les amplitudes sont toute égales à 5 :

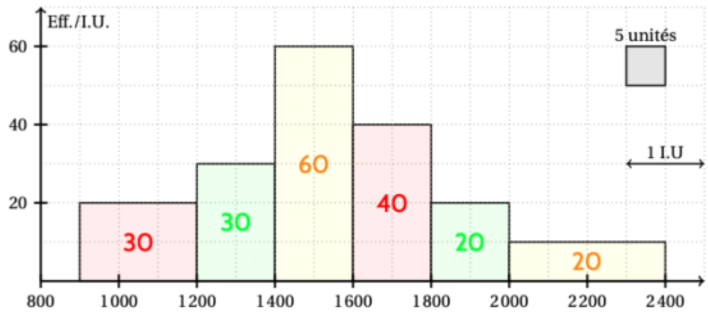


Illustration

Dans l'exemple suivant, les classes n'ont pas la même amplitude, on « s'arrange » donc pour créer une unité d'aire permettant de représenter la série.

- on choisit 200 pour unité de classe (généralement celui de la classe la plus petite) ;
- on détermine le nombre d'intervalles unitaires (I.U.) de chaque classe ;
- on détermine la hauteur des rectangles en divisant l'effectif de la classe par le nombre d'intervalles unitaires.

Valeurs	[900 ; 1200[	[1200 ; 1400[	[1400 ; 1600[	[1600 ; 1800[	[1800 ; 2000[	[2000 ; 2400[
Effectif	30	30	60	40	20	20
Largeur classe	300	200	200	200	200	400
Nb int. unit.	1,5	1	1	1	1	2
Eff. par int. unit.	20	30	60	40	20	10



Intro

Il existe également deux représentations « outils » qui permettent d'étudier plus finement une série :

- le diagramme en boîte à moustaches ;
- la courbe des ECC (ou des FCC).

Dans les deux cas de figure, on est sur une exploitation ou une recherche de paramètres statistiques (médiane, quartiles).

Rappel

Le polygone (ou la courbe) des effectifs cumulés (ou fréquences cumulées) est formé de segments de droite joignant les points :

- d'abscisses : la borne supérieure d'une classe pour le polygone des ECC ;
- d'ordonnées : l'effectif cumulé d'une classe ;
- partant du point d'ordonnée 0 et d'abscisse la borne inférieure de la première classe.

Exemple

On donne la série statistique suivante :

Valeurs x_i	[0;15[	[15;25[	[25;35[	[35;40[	[40;45[	[45;55[	[55;65[	[65;75[
Effectif n_i	15	20	50	30	35	25	15	10
Fréquence en %	7,5	10	25	15	17,5	12,5	7,5	5
FCC en %	7,5	17,5	42,5	57,5	75	87,5	95	100

Illustration