Dénombrement et statistiques

Table des matières

1 Dénombrement

Le but de ce paragraphe est de représenter et dénombrer une partie d'un ensemble donné (population, classe, objets,. . .), selon différents critères ou variables qui le composent (âge, sexe, activité professionnelle, . . .), en fonction des donnés fournies.

1.1 Deux variables indépendantes

Lorsque les donnés correspondants à ces deux variables ne dépendent pas l'une l'autre.

Exemple : Dans une classe de 34 élèves, 20 ont 16 ans, 25 pratiquent l'anglais en LV1 dont 13 élèves de 16 ans.

Analyse : les deux variables ici sont l'âge et la langue LV1.

Comment représenter ces informations ? Deux modèles sont possibles : un tableau double entrées ou un diagramme de Venn.

1.1.1 Tableau double entrées

On appelera :

- A : l'ensemble des élèves pratiquant l'anglais en LV1
  $\overline{A}$: l'ensemble des élèves ne pratiquant pas l'anglais en LV1
- B_{16} : l'ensemble des élèves de 16 ans
  $\overline{B_{16}}$: l'ensemble des élèves n'ayant pas 16 ans.

Remarque : On met une "barre" au dessus d'un ensemble pour exprimer son complémentaire, c'est à dire les éléments qui ne possèdent pas le critère de cet ensemble.

On obtient ainsi le tableau suivant en inscrivant les données fournies par l'énoncé.

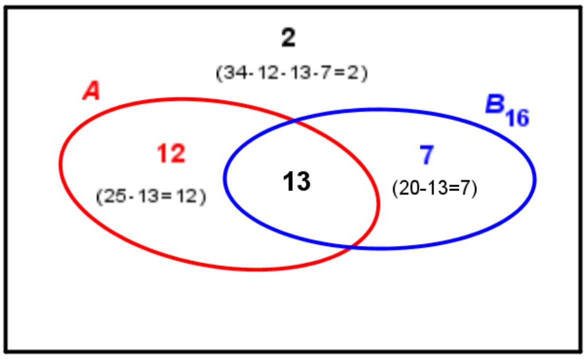
	A	$\overline{A}$	Total
B_{16}	13		25
$\overline{B_{16}}$			
Total	20		34

Par différence ou somme, on obtient le tableau complété suivant :

	A	$\overline{A}$	Total
B_{16}	13	12	25
$\overline{B_{16}}$	7	2	9
Total	20	14	34

1.1.2 Diagramme de Venn

L'autre possibilité consiste à faire des "patates" pour représenter la classe ainsi que ses différents critères. On obtient alors le diagramme suivant :



1.1.3 Application

Dans un groupe de 450 élèves, 30 % des élèves sont en Première, 64 % des élèves sont des filles et 75 filles sont en Première.

1 a) Traduire ces informations dans un tableau et compléter.

Comme on a des informations en valeurs absolues et en pourcentages, on traduira d'abord toutes ses informations en valeurs absolues.

$$30 \text{ \% de } 450 \quad 450 \times \frac{30}{100} = 135$$

$$64 \text{ \% de } 450 \quad 450 \times \frac{64}{100} = 288$$

On remplit alors un tableau que l'on complète par différence et somme.

Les deux variables ici sont la classe et le sexe.

On appelera :

- F : l'ensemble des filles.
  G : l'ensemble des garçons.
- P : L'ensemble des élèves de Première.
  $\overline{P}$: L'ensemble des élèves des autres classe.

	F	G	Total
P	75	60	135
$\overline{P}$	213	102	315
Total	288	162	450

b) Quelle est la part des filles dans les Première ?

Dans les 135 Premières, il y a 75 filles. Donc la part des filles dans les Premières est :

$$\frac{75}{135} \times 100 \simeq 56\%$$

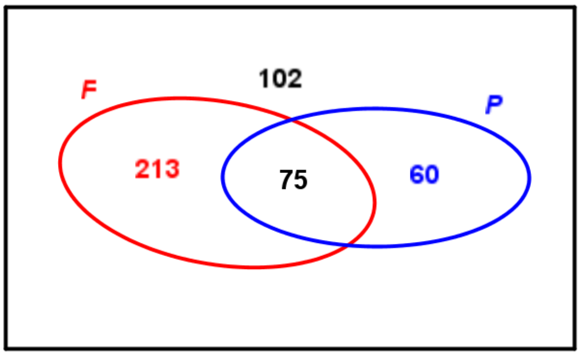
c) Quelle est la part des Premières parmi les filles ?

Dans les 288 filles, 75 sont en Première. Donc la part des Premières parmi les fille est :

$$\frac{75}{288} \times 100 \simeq 26\%$$

2 Faire un diagramme correspondant à ces deux critères.

On obtient alors :



1.2 Variables conditionnées

Lorsqu'une variable dépend d'une autre, on parle de variable conditionnée.

Exemple : Dans un lycée de 2 500 élèves, 38 % sont en classe de 2^{nde}, 28 % en 1^{re} et le reste en T^{le}. De plus, on sait que :

- 48 % des élèves de 2^{nde} sont externes.
- 65 % des élèves de 1^{re} sont externes.
- 52 % des élèves de T^{le} sont externes.

Quel est le pourcentage d'externe ?

Analyse : Les deux variables sont la classe et le statut (externe ou demi-pensionnaire) des élèves. D'après les données, on connaît de statut des élèves par classe. Le statut est donc conditionné à la classe.

Comment représenter ces informations ? Un arbre pondéré.

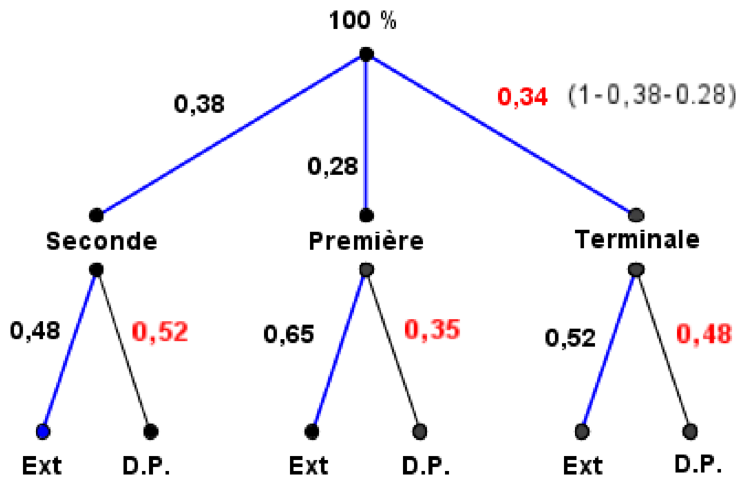
1.2.1 Arbre pondéré

Le premier niveau de l'arbre sera représenté par la classe (variable non conditionné) et le second par le statut.

Il convient d'énoncer les règles qui régissent un arbre pondéré :

- La loi des noeuds : la somme des coefficients autour d'un noeud est égal à 1.
- Lorsque l'on suit un chemin sur l'arbre, on multiplie les coefficients.
- Tous les coefficients sont exprimés par un nombre compris entre 0 et 1.

On obtient, en fonction des données, l'arbre suivant :



Pour déterminer le pourcentage d'externes, il faut tenir compte des trois chemins pour obtenir des externes. Le pourcentage d'externes est donc :

$$\begin{aligned}
\text{\% d'externes} &= 100 \times 0,38 \times 0,48 + 100 \times 0,28 \times 0,65 + 100 \times 0,34 \times 0,52 \\
&= 100(0,38 \times 0,48 + 0,28 \times 0,65 + 0,34 \times 0,52) \\
&= 100(0,1824 + 0,182 + 0,1768) \\
&= 54,12 \ \%
\end{aligned}$$

2 Statistiques

2.1 Objet

Sur une population (d'objets ou de personnes), on étudie un ou plusieurs critères ou variables. Les résultats obtenus constituent ce qu'on appelle une série statistique. Dans la suite du chapitre, on s'interessera aux séries d'une seule variable.

Pour un individu ou objet i, on associera la valeur de la variable x_i

$$i \longrightarrow x_i$$

L'ensemble des couples $(i; x_i)$ sera, dans la plupart des cas regroupés dans un tableau, qui constituera alors la série statistique.

Exemples :

- Sur un population d' élèves d'un classe, on étudie les notes obtenues en mathématiques.
- Sur une population de voitures, on étudie la couleur.
- Sur la population d'un pays, on étudie la taille des habitants de 18 ans ou plus.

Il existe plusieurs types de variables :

- Variable qualitative : la couleur par exemple. On ne peut quantifier la couleur. On représentera cette série avec un "camembert" par exemple. Ce ne sera pas l'objet de ce chapitre.
- Variable quantitative : on peut en distinguer de deux sortes : Variable discrète : qui ne peuvent prendre que peu de valeurs possibles (le nombre d'enfants par foyer par exemple). On représentera cette série avec un diagramme à bâtons.
  Variable continue : Qui peuvent prendre autant de valeurs que l'on souhaite (la taille d'un adulte par exemple). Dans la pratique, on ne sélectionnera qu'un dizaine de catégories réparties par classe. Ceci dans un souci d'analyse de la série. On représentera cette série dans un histogramme.

2.2 Paramètres de position

Pour étudier un série statistique, nous avons besoin d'outil. Un de ceux-ci est le paramètre de position : où se situe le milieu de la série. On pense, bien évidement à la moyenne, mais on peut se doter d'une autre sorte de milieu : la médiane.

2.2.1 La moyenne

1) La moyenne simple :

Si la série ne comporte qu'un petit nombre de données. On somme les x_i et l'on divise par le nombre de donné N. On note $\bar{x}$ la moyenne obtenue. On a alors la formule suivante :

$$\bar{x} = \frac{\sum x_i}{N}$$

Exemple : soit les cinq notes de mathématiques suivantes :

$$8 \;\; ; \;\; 12 \;\; ; \;\; 9,5 \;\; ; \;\; 17 \;\; ; \;\; 13$$

Leur moyenne est alors :

$$\bar{x} = \frac{8 + 12 + 9,5 + 17 + 13}{5} = \frac{59,5}{5} = 11,9$$

2) La moyenne pondérée :

Lorsque le nombre de données est plus important, on est amené à remplir un tableau d'effectifs. On note alors x_i une valeur prise par la valiable et n_i son effectif. N étant toujours le nombre total de données, on a alors :

$$\bar{x} = \frac{\sum n_i \times x_i}{N}$$

Exemple : Soit les notes de mathématiques obtenues par les 36 élèves d'une classe de 1re L :

Notes (x_i)	8	9	10	11	12	13	14
Effectifs (n_i)	6	2	7	3	4	8	6

On a alors, la moyenne de la classe suivante :

$$\bar{x} = \frac{8 \times 6 + 9 \times 2 + 10 \times 7 + 11 \times 3 + 12 \times 4 + 13 \times 8 + 14 \times 6}{36} = \frac{405}{36} = 11,25$$

3) Moyenne de deux séries statisques

Lorsque deux séries 1 et 2 ont pour moyenne respective $\bar{x}_1$ et $\bar{x}_2$ et comme effectif respectif n_1 et n_2, la moyenne des deux séries $\bar{x}_T$ est égale à :

$$\bar{x}_T = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

Exemple : Dans une entreprise de 60 salariés, le salaire moyen des hommes est de 1 500 € net et le salaire moyen des femmes de 1 300 € net. Sachant qu'il y a 42 femmes dans l'entreprise, quel est le salaire net moyen des salariés ?

S'il y a 42 femmes, il y a : $60 - 42 = 18$ hommes. Le salaire net moyen des salariés en euros est égal à :

$$\bar{x}_T = \frac{18 \times 1\ \ 500 + 42 \times 1\ \ 300}{60} = \frac{81\ \ 600}{60} = 1360$$

2.2.2 La médiane

On cherche ici à séparer la série en deux effectifs égaux.

Définition 1 :

On appelle médiane d'une série ordonnée, la valeur Me qui partage cette série en deux effectifs égaux.

Deux cas peuvent se présenter :

1. Le nombre de données est impair. Le nombre $\dfrac{N+1}{2}$ est alors un nombre entier. On prendra alors la valeur correspondante dans la série.

   Soit la série de notes suivante :

   $$8 \ ; \ 12 \ ; \ 9,5 \ ; \ 13 \ ; \ 17$$

   On ordonne la série dans l'ordre croissant, on obtient alors :

   $$8 \ ; \ 9,5 \ ; \ 12 \ ; \ 13 \ ; \ 17$$

   On calcule :

   $$\frac{N+1}{2} = \frac{5+1}{2} = 3$$

   On prend la troisième valeur de la série :

   $$Me = 12$$

2. Le nombre de données est pair. Le nombre $\dfrac{N+1}{2}$ n'est pas entier, il est compris entre deux entiers. On prendra alors le milieu des valeurs correspondantes.

   Soit la série de notes suivante :

   $$8 \ ; \ 9,5 \ ; \ 11 \ ; \ 12 \ ; \ 13 \ ; \ 17$$

   On calcule :

   $$\frac{N+1}{2} = \frac{6+1}{2} = 3,5$$

   On prend le milieu de la troisième et quatrième valeur de la série :

   $$Me = \frac{11+12}{2} = 11,5$$

2.2.3 Quartiles

On peut, comme pour la médiane, définir deux autres paramètres de position : le premier et troisième quartile

Définition 2 :

Le premier quartile Q_1 d'une série ordonnée est la plus petite valeur pour laquelle 25 % au moins des valeurs de la série sont égales ou inférieures à celle-ci.

Le troisième quartile Q_3 est la plus petite valeur pour laquelle 75 % au moins des valeurs de la série sont égale ou inférieures à celle-ci.

On appelle l'intervalle interquartile, l'intervalle : $IQ = [Q_1; Q_3]$

L'écart interquartile est alors : $e = Q_3 - Q_3$

Dans la pratique, on calcule les quantités : $\dfrac{N}{4}$ et $\dfrac{3N}{4}$ en prenant la valeur immédiatement au dessus.

Exemple : On connaît la taille (en cm) d'un groupe de 45 enfants de 5 à 7 ans. On obtient alors la série :

106	109	110	111	113	114	116	118	121
107	109	111	111	114	114	117	120	121
108	109	111	112	114	115	117	120	121
108	109	111	112	114	116	118	120	123
109	110	111	113	114	116	118	121	126

On calcule :

$$\frac{N}{4} = \frac{45}{4} = 11,25 \quad \text{et} \quad \frac{3N}{4} = \frac{135}{4} = 33,75$$

On prend donc la 12e valeur et la 34e valeur pour le 1er et 3e quartile.

$$Q_1 = 111 \qquad \text{et} \qquad Q_3 = 118$$

On obtient donc l'intervalle interquartile : $IQ = [111, 118]$

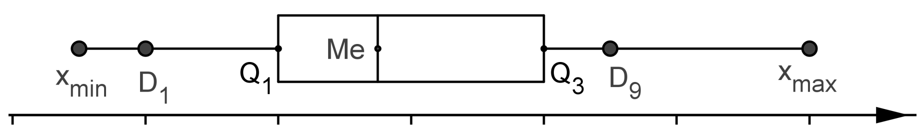
L'écart interquartile est : $e = 118 - 111 = 7$

2.2.4 Diagramme en boîte

Pour résumer les différentes valeurs que l'on a déterminées, on réalise un diagramme appelé "diagramme en boîte". Dans ce diagramme figure : les valeurs extrêmes : valeurs minimum et maximum, les quartiles et la médiane.

Remarque : Lorsque la série a beaucoup de valeurs, on peut être amené à diviser la série en 10 parties égales : ce sont les déciles. Les valeurs de premier décile D_1 et du neuvième décile D_9 remplace alors les valeurs extrêmes de la serie dans le diagramme en boîte.

On a alors :



Exemple : Reprenons l'exemple de la taille des 45 enfants.

On détermine la médiane :

$\dfrac{N+1}{2} = \dfrac{45+1}{2} = 23$. On prend la 23e valeur de la série :

$$Me = 114$$

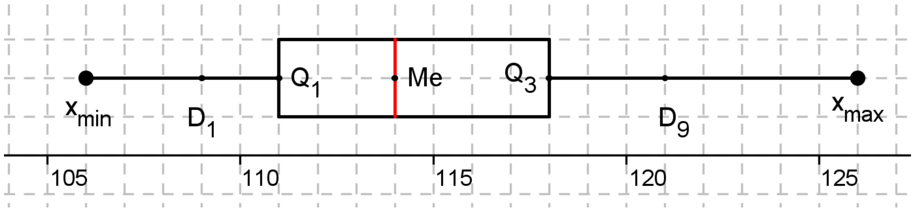
Les valeurs extrèmes sont respectivement : 106 et 126.

On peut éventuellement calculer les 1er et 9e déciles. On calcule alors : $\frac{45}{10} = 4,5$ et $9 \times \frac{45}{10} = 40,5$

On prend respectivement la 5e et la 41e valeur, on obtient alors :

$$D_1 = 109 \qquad \text{et} \qquad D_9 = 121$$

On otient alors le diagramme en boîte suivant :



Pour étudier la série, on peut analyser :

1. La médiane $Me = 114$
2. L'écart interquartile qui correspond à 50 % de l'effectif autour de la médiane. Ici $e = 7$
3. L'étendue de la série : $x_{max} - x_{min} = 126 - 106 = 20$

2.3 Paramètres de dispersion

2.3.1 Variance et écart type

Définition 3 :

Dans une série de N valeurs et de moyenne $\bar{x}$, on appelle variance V, la valeur qui correspond à la moyenne des écarts au carré par rapport à la moyenne. On a donc suivant que la série est simple ou pondérée :

$$V = \frac{\sum (x_i - \bar{x})^2}{N} \qquad \text{ou} \qquad V = \frac{\sum n_i \times (x_i - \bar{x})^2}{N}$$

L'écart type σ représente alors la racine carrée de la variance, soit :

$$\sigma = \sqrt{V}$$

Exemples : On s'interesse aux notes de mathématiques des élèves Coraline et Séverine. Les six notes obtenues sont consignées dans le tableau suivant :

Coraline	12	8	5	16	9	10
Séverine	10	11	12	10	8	9

On calcule d'abord la moyenne pour chque élève :

$$\overline{x_c} = \frac{12+8+5+16+9+10}{6}\frac{60}{6} = 10 \quad \text{et} \quad \overline{x_s} = \frac{10+11+12+10+8+9}{6} = \frac{60}{6} = 10$$

On calcule ensuite les variances pour chaque élève :

$$\begin{aligned} V_c &= \frac{(12-10)^2+(8-10)^2+(5-10)^2+(16-10)^2+(9-10)^2+(10-10)^2}{6} \\ &= \frac{4+4+25+36+1+0}{6} = \frac{70}{6} \simeq 11,67 \end{aligned}$$

$$\begin{aligned} V_s &= \frac{(10-10)^2+(11-10)^2+(12-10)^2+(10-10)^2+(8-10)^2+(9-10)^2}{6} \\ &= \frac{0+1+4+0+4+1}{6} = \frac{10}{6} \simeq 1,67 \end{aligned}$$

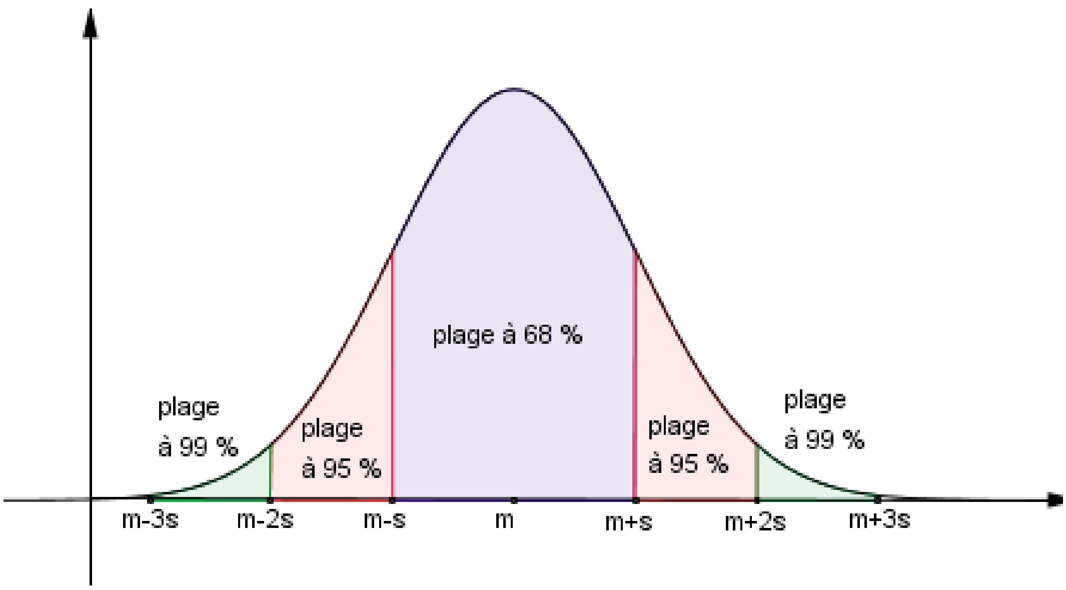
On a alors les écart types suivants :

$$\sigma_c = \sqrt{V_c} \simeq \sqrt{11,67} \simeq 3,4$$
$$\sigma_s = \sqrt{V_s} \simeq \sqrt{1,67} \simeq 1,3$$

2.3.2 Modèle gaussien

Définition 4 : Le modèle gaussien, représente une série symétrique par rapport à la moyenne (dans ce cas $\overline{x} = Me$) où l'on définit des plages de normalité ou intervalle de confiance.

On a donc le schéma suivant :



m représente la moyenne et s l'écart type

Les plages de normalité sont définies comme ci-dessous :

- Plage de normalité à 68 % : $[m - s; m + s]$
- Plage de normalité à 95 % : $[m - 2s; m + 2s]$
- Plage de normalité à 99 % : $[m - 3s; m + 3s]$