جامعة أبو بكر بلقايد
UNIVERSITY OF TLEMCEN
كلية علوم الطبيعة والحياة علوم الارض والكون

جامعة أبي بكر بلقايد - تلمسان
UNIVERSITY OF BELKAID ABOU BEKR – TLEMCEN
كلية العلوم
FACULTY OF SCIENCES
قسم الرياضيات
DEPARTMENT OF MATHEMATICS

Cours de Biostatistique

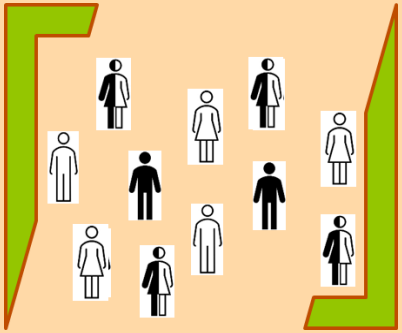
Licence Biologie moléculaire

Année universitaire 2022/2023

Joanna Dib, Djilali Ameur, Majda Dali-Sahi

Une étude statistique comprend en général les étapes suivantes :

- Procéder à une enquête par sondage ou recensement et collecte des données
- Précision du caractère de la variable statistique étudiée
- Présentation des données dans un tableau
- Représentation de cette série statistique à l'aide d'un diagramme ou histogramme
- Calcul des paramètres permettant de caractériser toute la série statistique à l'aide de quelques nombres tels que la moyenne, la variance, l'écart–type…



CHAPITRE I

TERMINOLOGIES STATISTIQUES

I. Vocabulaires statistiques

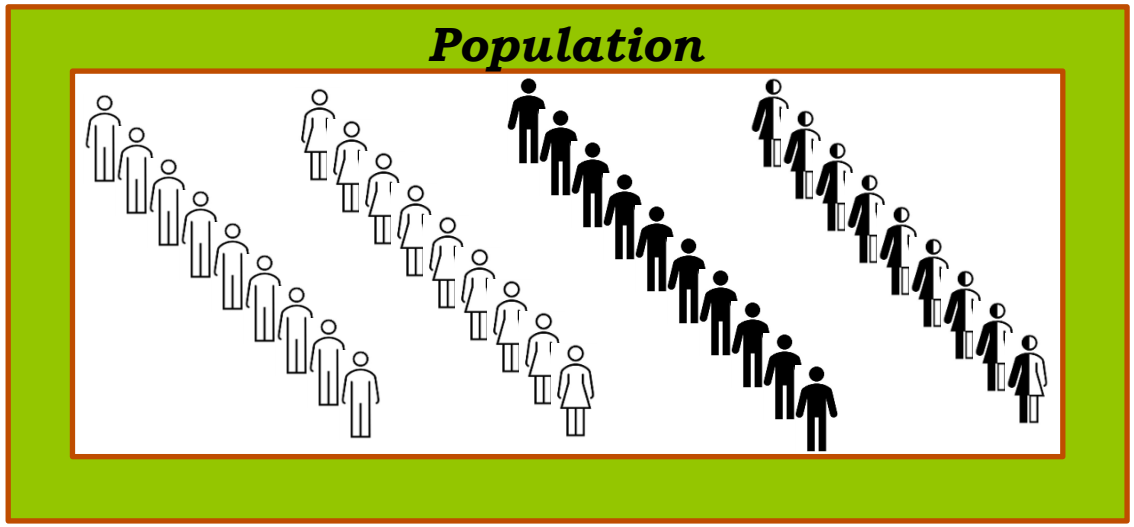
Nous parlons de ***recensement*** lorsque l'on fait une étude exhaustive d'une population

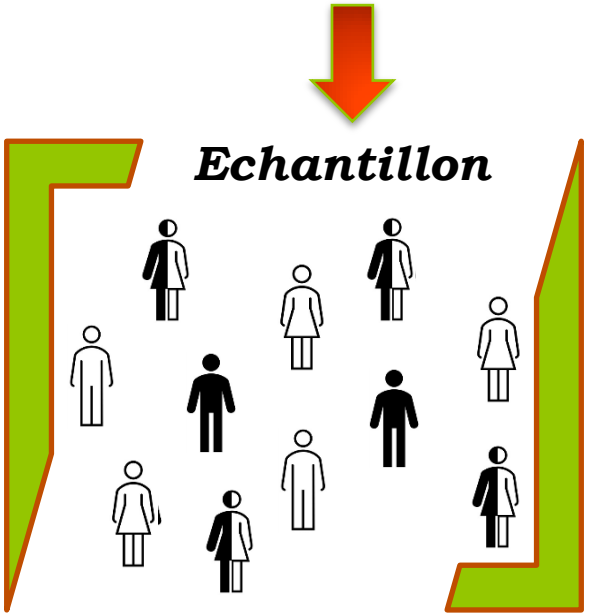
Nous nous intéressons à un ensemble fini dit ***population*** que l'on observe et qui sera soumis à une analyse statistique

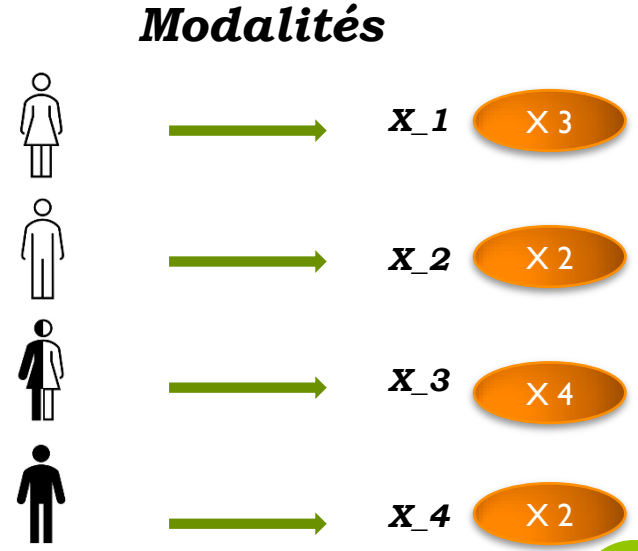
Chaque élément de cet ensemble est une ***unité statistique*** ou ***individu***. Nous supposons que la variable prend toujours une seule valeur sur chaque unité

Sur un sous-ensemble de la population considérée dite ***échantillon***, nous nous proposons d'observer un phénomène : un ***caractère*** ou une ***variable statistique*.**
Nous supposons que la variable prend toujours une seule valeur sur chaque unité. Les variables sont désignés par X, Y, Z

Les valeurs possibles de la variable sont appelées ***modalités***. Nous notons les modalités en utilisant la même lettre que le caractère, mais en minuscule et indicée. Ainsi nous notons x_i la i-ème modalité du caractère X et y_j la j-ème modalité du caractère Y


Population


Echantillon


Modalités
X_1
X 3
X_2
X 2
X_3
X 4
X_4
X 2

II. Typologie de caractère

Un caractère étudié peut être soit ***qualitatif***, soit ***quantitatif***

II.1 Variable statistique qualitative

C'est un caractère qui ne peut être mesuré ni repéré par un nombre

Les modalités du caractère qualitatif rangent les unités de la population étudiée en catégories

Le caractère qualitatif peut être ***ordinal*** s'il on peut ordonner les modalités de la variable statistique étudié ou ***nominal*** dans le cas contraire

Exemple 1 : *Variable qualitative nominale*

La couleur des yeux est un caractère héréditaire influencé par plus d'un gène qui permettent d'expliquer les trois grands types de couleurs phénotypiques des yeux chez l'être humain : Brun, vert et bleu.

Nous nous intéressons à la ***variable statistique*** "couleur des yeux" notée X et à la ***série statistique*** des valeurs prises par X sur 20 personnes. La codification est :

Br : brun
Ve : vert
Bl : bleu

Le ***domaine de la variable*** X est {Br, Ve, Bl}. Considérons la série statistique suivante :

Ve	Br	Br	Bl	Br	Ve	Br	Br	Br	Ve
Bl	Br	Br	Ve	Br	Br	Bl	Ve	Br	Br

Ici N =20

$$x_1 = Ve,\ x_2 = Br,\ x_3 = Br,\ x_4 = Bl,\ x_5 = Br,\ ...,\ x_{20} = Br.$$

Exemple 2 : *Variable qualitative nominale*

Une enquête effectuée auprès de L'Atlas du Diabète de la FID (9ème Édition 2019) selon le Top 10 des pays ou territoires en nombre d'adultes (20 à 79 ans) vivant avec le diabète en 2019 **en millions**, se répartissent de la manière suivante :

Pays ou territoire	Nombre de personnes vivant avec le diabète
Chine	116,4
Inde	77,0
États-Unis	31,0
Pakistan	19,4
Brésil	16,8
Mexique	12,8
Indonésie	10,7
Allemagne	9,5
Égypte	8,9
Bangladesh	8,4
Total	310,9

Exemple 3 : *Variable qualitative ordinale*

Nous considérons la variable statistique "niveau universitaire" des étudiants, dont le domaine est : {L1, L2, L3, M1, M2, Doctorat, Post Doctorat, ...}

Exemple 4 : *Variable qualitative ordinale*

Une enquête effectuée auprès du ministère de l'Education Nationale Algérienne Direction Technique chargée des Statistiques Régionales, de l'agriculture et de la Cartographie Direction des publications et de la Diffusion, N871 selon les principaux agrégats de l'année scolaire 2018-2019, se répartissent de la manière suivante

Niveau d'enseignement	Elèves	Enseignants
Préparatoire	495.481	17.791
Primaire	4.513.749	199.850
Moyen	2.979.737	159.065
Secondaire	1.222.673	102.279
Total	9.211.640	478.985

II.2 Variable statistique quantitative

C'est un caractère qui peut être mesuré et repéré par un nombre.

La variable peut être de deux natures :

Variable statistique quantitative discrète

- Si elle ne peut prendre que des valeurs isolées ou de modalités possibles, souvent entières, dans l'intervalle où elle varie.

Variable statistique quantitative continue

- Si elle peut prendre n'importe quelle valeur dans l'intervalle, même s'il ne prend pas effectivement toutes ces valeurs.

Exemple 5 : *Variable quantitative discrète*

Un quartier est composé de 40 ménages et la ***variable*** Y représente le nombre d'enfants par ménage. Les valeurs de la variable sont :

0	0	0	0	1	1	1	1	1	2
2	2	2	2	2	2	2	3	3	3
3	3	3	4	4	4	4	4	4	4
5	5	5	5	5	5	6	6	6	8

Le ***domaine de la variable*** Y est {0,1,2,3,4,5,6,8}. Ici N=40

$$x_1 = 0,\ x_2 = 0,\ x_3 = 0,\ x_4 = 0,\ x_5 = 1,\ \ldots,\ x_{20} = 8.$$

Exemple 6 : *Variable quantitative continue*

On mesure la taille en centimètres de 50 étudiants d'une classe :

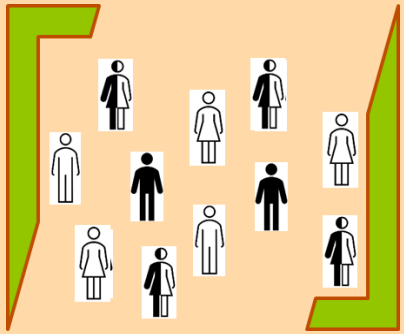
151	152	152	152	152
153	153	154	154	154
155	155	155	155	155
156	156	157	157	157
158	158	159	160	160
160	161	161	161	161
162	163	163	163	164
165	166	166	167	167
167	168	168	169	170
170	172	172	174	165

On a ***les classes*** des tailles définies préalablement comme il suit :

$[150; 155[$
$[155; 160[$
$[160; 165[$
$[165; 170[$
$[170; 175[$

Le ***domaine de la variable*** Z est { [150;155[, [155;160[, [160;165[, [165;170[, [170;175[}.

Ici N=50



CHAPITRE II

STATISTIQUE DESCRIPTIVE UNIDIMENSIONNELLE

I. Représentation des données statistiques

Il existe plusieurs moyens de description statistique :

- Soit la présentation brute des données.
- Soit des présentations par tableaux numériques.
- Soit des représentations graphiques.
- Soit des résumés numériques fournis par un petit nombre de paramètres caractéristiques dites de position ou de dispersion.

I.1 Tableau statistique

- Une ***distribution statistique*** donne en fonction de chaque modalité du caractère le nombre d'individus de la population prenant cette modalité.
- Elle est présentée sous la forme d'un tableau à une entrée ou à un caractère qui est un mode synthétique de présentation des données :

Modalité x_i	**Effectif** n_i	**Fréquence** f_i
x_1	n_1	f_1
x_2	n_2	f_2
$\vdots$	$\vdots$	$\vdots$
x_i	n_i	f_i
$\vdots$	$\vdots$	$\vdots$
x_r	n_r	f_r
Total	N	1

I.2 Effectif et effectif cumulé

- ***L'effectif total*** est le nombre d'individus d'une population, il est noté N.
- ***L'effectif de chaque modalité*** x_i d'un caractère X, noté n_i, correspond au nombre d'individus présentant cette modalité

Effectif d'une modalité dans une observation statistique = nombre de fois que cette modalité du caractère apparaît dans la population étudiée

Nous avons :

$$\sum_i n_i = N$$

La somme des effectifs de l'ensemble des modalités dans une observation statistique = l'effectif total

- ***L'effectif cumulé d'une modalité*** x_i d'un caractère X correspond au nombre d'individus présentant au plus cette modalité. Nous pourrons distinguer deux cas :

- ***L'effectif cumulé croissant***, noté $n\_i \nearrow$ associé à une valeur qui est la somme des effectifs des valeurs inférieures. Elle consiste à cumuler successivement, par ordre croissant, les effectifs à partir des plus faibles valeurs en ajoutant à chaque fois la fréquence suivante.
- ***L'effectif cumulé décroissant***, noté $n\_i \swarrow$, associé à une valeur qui est la somme des effectifs des valeurs supérieures. C'est la sommation successive, par ordre décroissant, des effectifs en commençant par les plus grandes valeurs.

Exemple 6 (suite) : *Effectif cumulé*

Reprenons l'exemple de la variable Z portant sur la taille en centimètre de 50 étudiants et complétons le tableau statistique associé avec les colonnes des effectifs cumulés croissants et décroissants :

Modalité z_i	Effectif n_i	Effectif cumulé $\nearrow$ $n_i \nearrow$	Effectif cumulé $\swarrow$ $n_i \swarrow$
$[150; 155[$	10	10	50
$[155; 160[$	13	23	40
$[160; 165[$	12	35	27
$[165; 170[$	9	44	15
$[170; 175[$	6	50	6
Total	N=50		

I.3 Fréquence et fréquence cumulée

- La ***fréquence de chaque modalité*** x_i d'un caractère X, noté f_i, correspond à la proportion d'individus de la population présentant cette modalité. La fréquence d'une modalité peut s'exprimer par un nombre décimal inférieur ou égal à 1 et elle peut aussi s'exprimer en pourcentage. Elles est égale à :

$$f_i = \frac{n_i}{N} \quad avec \quad \sum_i f_i = 1$$

La somme des fréquences de l'ensemble des modalités dans une observation statistique = 1

- ***La fréquence cumulée croissante***, noté $f\_i\nearrow$ associée à une valeur qui est la somme des fréquences des valeurs inférieures. Elle consiste à cumuler successivement, par ordre croissant, les fréquences à partir des plus faibles valeurs en ajoutant à chaque fois la fréquence suivante.
- ***La fréquence cumulée décroissante***, noté $f\_i\swarrow$ associée à une valeur qui est la somme des fréquences des valeurs supérieures. C'est la sommation successive, par ordre décroissant, des fréquences en commençant par les plus grandes valeurs.

Exemple 6 (suite) : *Fréquence cumulée*

Reprenons l'exemple de la variable Z portant sur la taille en centimètre de 50 étudiants et complétons le tableau statistique associé avec les colonnes des effectifs cumulés croissants et décroissants, et celles des fréquences cumulées croissantes et décroissantes :

Modalité z_i	Effectif n_i	Effectif cumulé $\nearrow$ $n_i \nearrow$	Effectif cumulé $\swarrow$ $n_i \swarrow$	Fréquence f_i	Fréquence cumulé $\nearrow$ $f_i \nearrow$	Fréquence cumulé $\swarrow$ $f_i \swarrow$
[150; 155[	10	10	50	0.20	0.20	1.00
[155; 160[	13	23	40	0.26	0.46	0.80
[160; 165[	12	35	27	0.24	0.70	0.54
[165; 170[	9	44	15	0.18	0.88	0.30
[170; 175[	6	50	6	0.12	1.00	0.12
Total	N=50			1		

Ligne 4

L'interprétation de la ligne 4 nous donne :

- 9 étudiants, ce qui correspond à 18% de la population étudiée, ont une taille comprise entre 165 et 170 cm
- 44 étudiants, ce qui correspond à 88% de la population étudiée, ont une taille inférieure à 170 cm
- 15 étudiants, ce qui correspond à 30% de la population étudiée, ont une taille supérieure à 165 cm.

I.4 Représentations graphiques

I.4.1 Variable statistique qualitative

Pour représenter une variable statistique qualitative, on se sert de trois sortes de représentations graphiques :

- Les diagrammes en bâtons verticaux ou horizontaux.
- Les graphiques ou diagrammes en barres.
- Les diagrammes circulaires.

➢ Diagramme en bâtons vertical

- Le diagramme en bâtons est utilisé pour représenter les séries statistiques correspondant à un caractère discret à l'aide de ***segments*.**

- C'est un graphique dans lequel les modalités de la variable statistique étudiée sont représentées sur l'axe horizontal des x, un axe non gradué, et les effectifs correspondants sur l'axe vertical des y, un axe gradué.

- À chaque modalité correspond un bâton d'une longueur proportionnelle aux effectifs n_i ou à la fréquence f_i selon qu'il s'agit d'un diagramme des effectifs ou d'un diagramme des fréquences puisque les fréquences présentent le même profil que les effectifs.

Exemple 1 (suite) : *Diagramme en bâtons vertical*

Reprenons l'exemple des couleur des yeux , nous obtenons donc les graphiques suivants

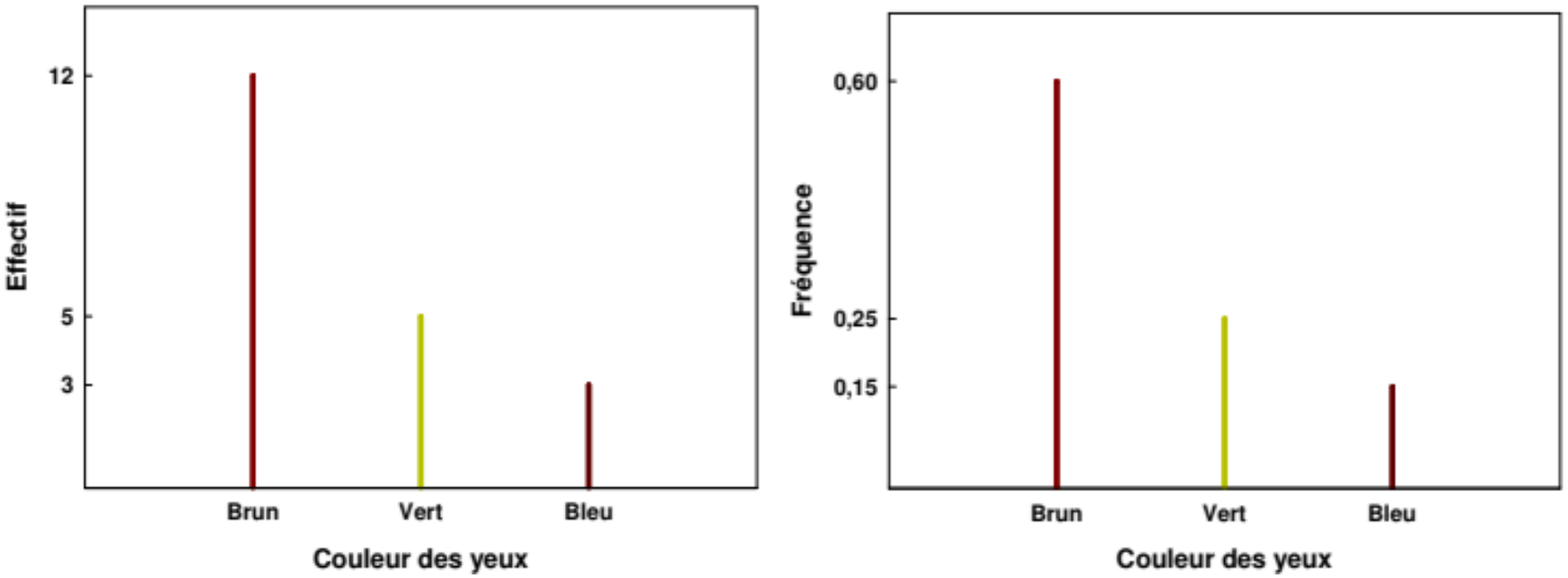


Figure 1 : Diagramme en bâtons vertical des effectifs et des fréquences de la variable couleur des yeux

➢ Diagramme en bâtons horizontal

Pour un diagramme en bâtons horizontal, c'est la largeur du rectangle qui représente la valeur de la variable continue et la hauteur de ce rectangle n'a aucune interprétation statistique.

Exemple 1 (suite) : *Diagramme en bâtons horizontal*

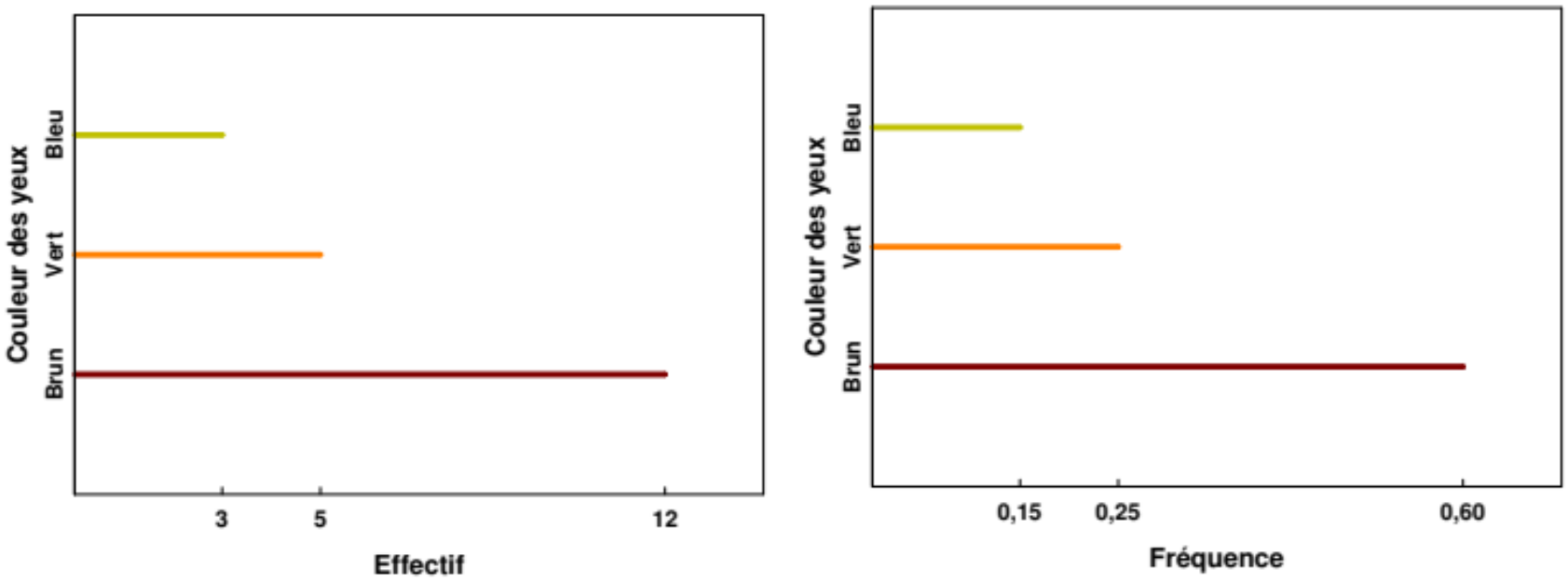


Figure 2 : Diagramme en bâtons horizontal des effectifs et des fréquences de la variable couleur des yeux

➢ Graphique ou diagramme en barres

- Un diagramme en barres, relativement fréquent, est une représentation graphique de données statistiques à l'aide de rectangles de même largeur.
- Les modalités du caractère étudié sont représentées sur l'axe horizontal et les effectifs respectifs sur l'axe vertical. À chaque valeur correspond une barre à la place des bâtons.
- Les hauteurs des barres sont proportionnelles aux effectifs représentés.
- Dans le graphique en barres, on ne doit pas joindre les rectangles qui doivent rester espacés et distincts, vu que le caractère est discontinu, et que les rectangles sont de même largeur.
- Nous pourrons se servir aussi des graphiques à barres empilées composés de barres dont les éléments d'une même série sont placés l'un à la suite de l'autre à l'horizontale. La longueur totale de chaque barre est la somme des éléments de la catégorie.

Exemple 1 (suite) : *Diagramme en barres*

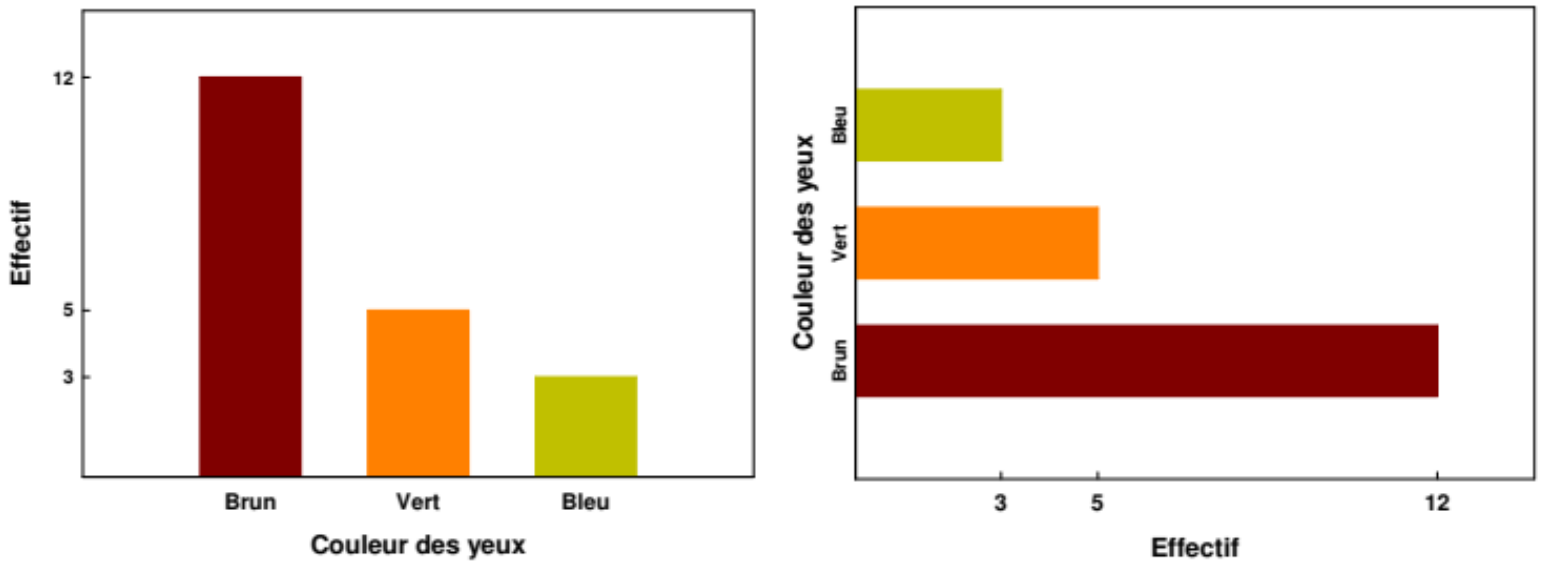


Figure 3 : Diagramme en barres vertical et horizontal des effectifs de la variable couleur des yeux

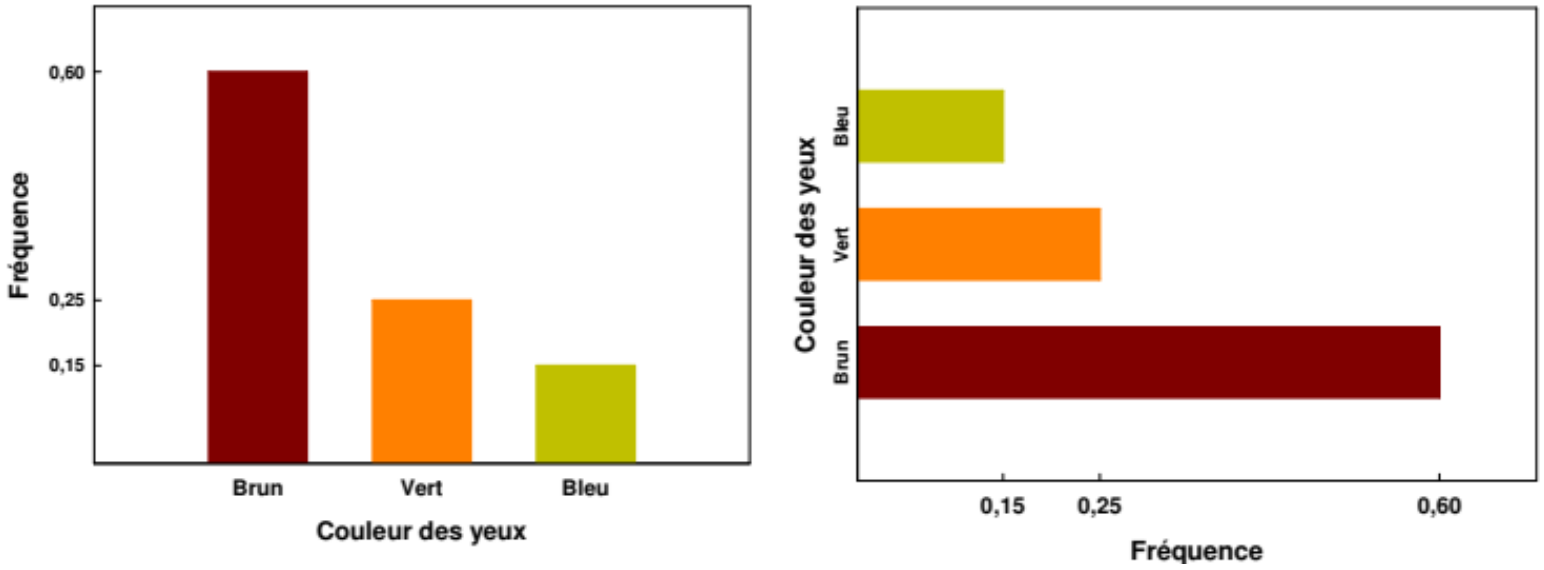


Figure 4 : Diagramme en barres vertical et horizontal des fréquences de la variable couleur des yeux

➢ Diagramme circulaire

- Un diagramme circulaire, communément appelé ***Camembert*** est une représentation graphique de données statistiques sous la forme d'un disque partagé en secteurs circulaires proportionnels.
- Pour le camembert, c'est la surface allouée à la modalité qui est proportionnelle à la fréquence. La part correspondant de chaque modalité x_i correspond à un angle α au centre égal en degrés à :

$$\alpha_i = \frac{n_i}{N} \times 360^0 = f_i \times 360^0$$

où n_i et f_i sont l'effectif et la fréquence de la modalité x_i de la variable statistique X.

- Le diagramme circulaire peut être aussi un ***donut chart*** qui est un camembert troué au milieu. Dans ce cas, c'est la longueur de l'arc de cercle correspondant à chaque catégorie qui représente la part de chaque catégorie dans le tout représenté.

Exemple 1 (suite) : *Diagrammes circulaires*

Reprenons l'exemple des couleur des yeux, les mesures des secteurs sont proportionnelles aux effectifs représentés, le coefficient de proportionnalité étant ici égal à $\frac{360}{20}$

Modalité x_i	Effectif n_i	Fréquence f_i	Mesure du secteur en degré A^0
Br	12	0.6	216
Ve	5	0.25	90
Bl	3	0.15	54
Total	N=20	1	360

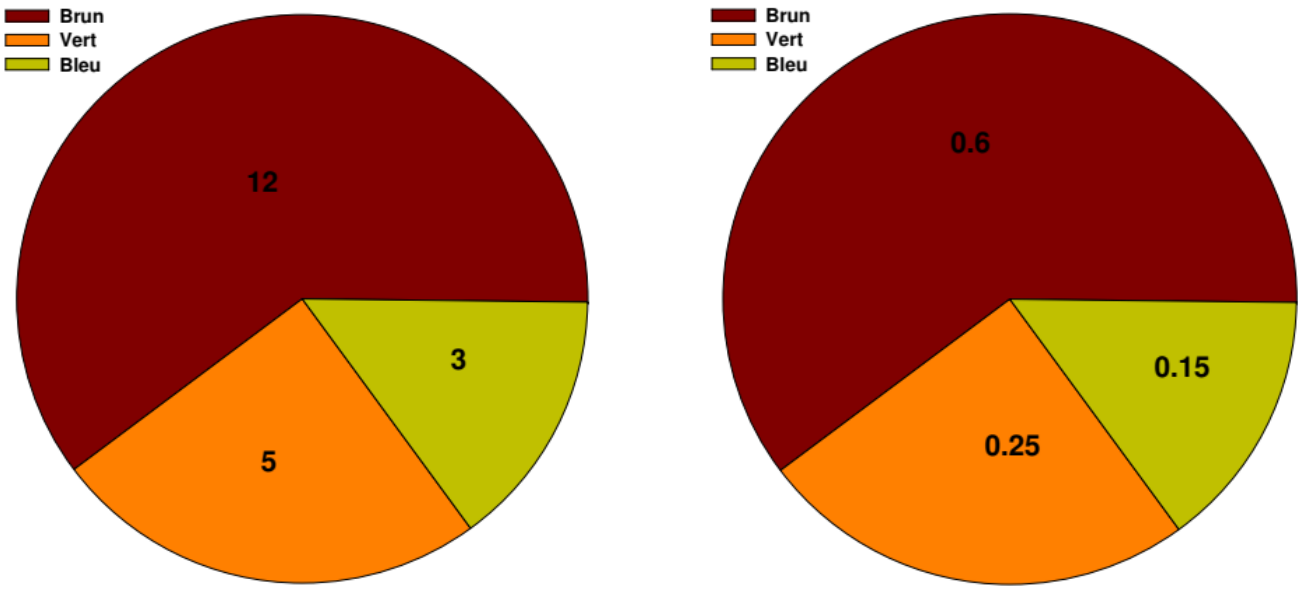


Figure 5 : Diagrammes circulaires des effectifs et des fréquences de la variable couleur des yeux

I.4.2 Variable statistique quantitative discrète

- Pour des variables qualitatives, nous privilégions les diagrammes circulaires dits "en camembert", demi-circulaire ou rectangulaire.
- Nous pourrons utilisés aussi les mêmes diagrammes évoqués ci-dessus.
- On ajoute à ceci le diagramme des effectifs et des fréquences cumulés croissants et décroissants.

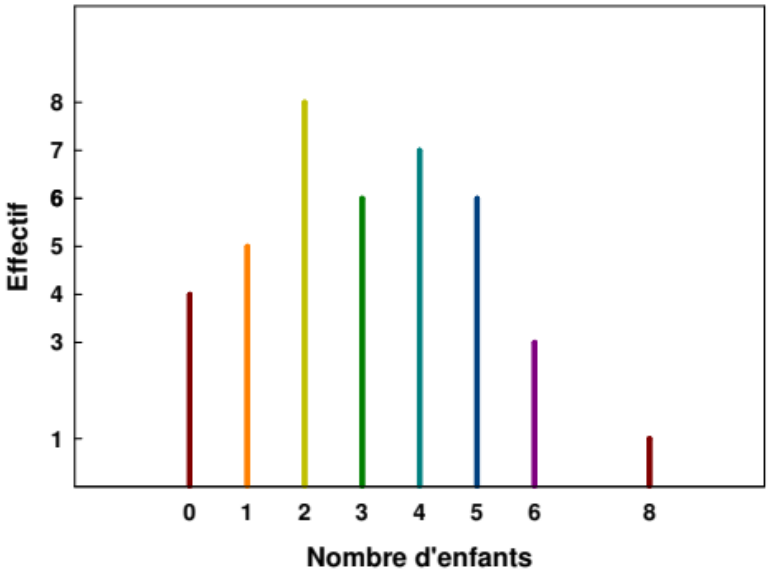
➢ Diagramme des effectifs et des fréquences cumulés croissants et décroissants

C'est un diagramme en bâtons dont la hauteur correspond aux effectifs ou fréquences cumulés croissants ou décroissants.

Exemple 5 (suite) : Représentations graphiques

Reprenons l'exemple du nombre d'enfants par ménage (Variable statistique *Y*), complétons le tableau statistique associé avec les colonnes des effectifs cumulés croissants et décroissants, et celles des fréquences cumulées croissantes et décroissantes :

Modalité y_i	**Effectif** n_i	**Effectif cumulé** $\nearrow$ $n_i \nearrow$	**Effectif cumulé** $\swarrow$ $n_i \swarrow$	**Fréquence** f_i	**Fréquence cumulée** $\nearrow$ $f_i \nearrow$	**Fréquence cumulée** $\swarrow$ $f_i \swarrow$
0	4	4	40	0.100	0.100	1
1	5	9	36	0.125	0.225	0.900
2	8	17	31	0.200	0.425	0.775
3	6	23	23	0.150	0.575	0.575
4	7	30	17	0.175	0.750	0.425
5	6	36	10	0.150	0.900	0.250
6	3	39	4	0.075	0.975	0.100
8	1	40	1	0.025	1	0.025
Total	N=40			1		



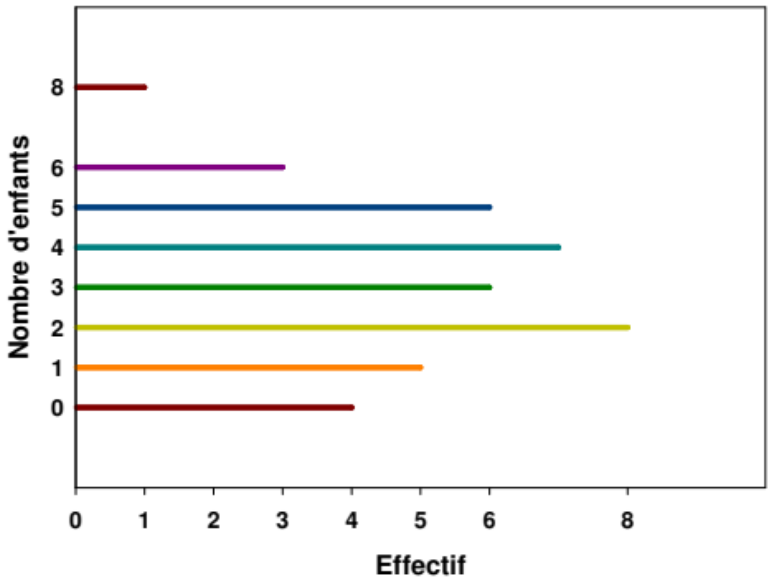
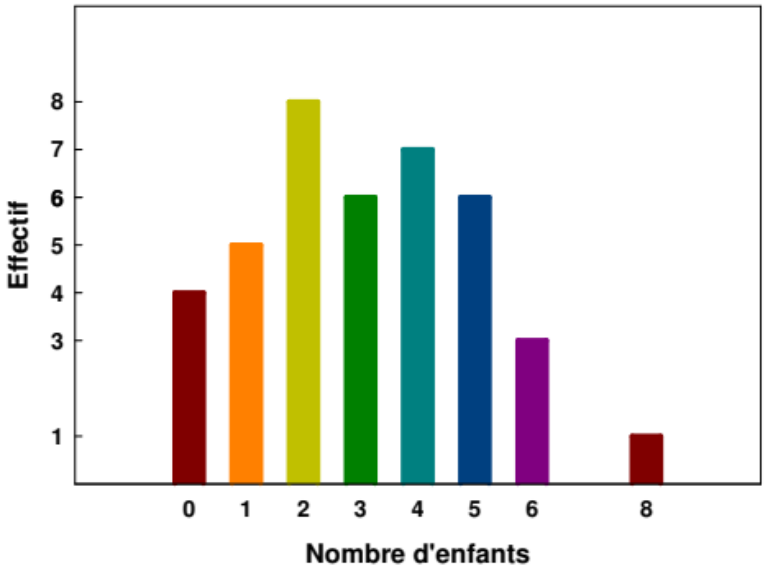


Figure 6 : Diagramme en bâtons vertical et horizontal des effectifs



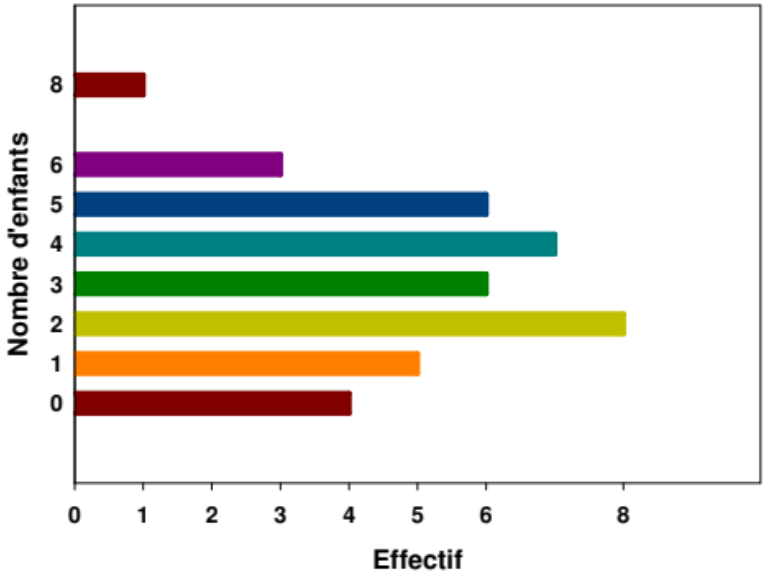


Figure 7 : Diagramme en barres vertical et horizontal des effectifs

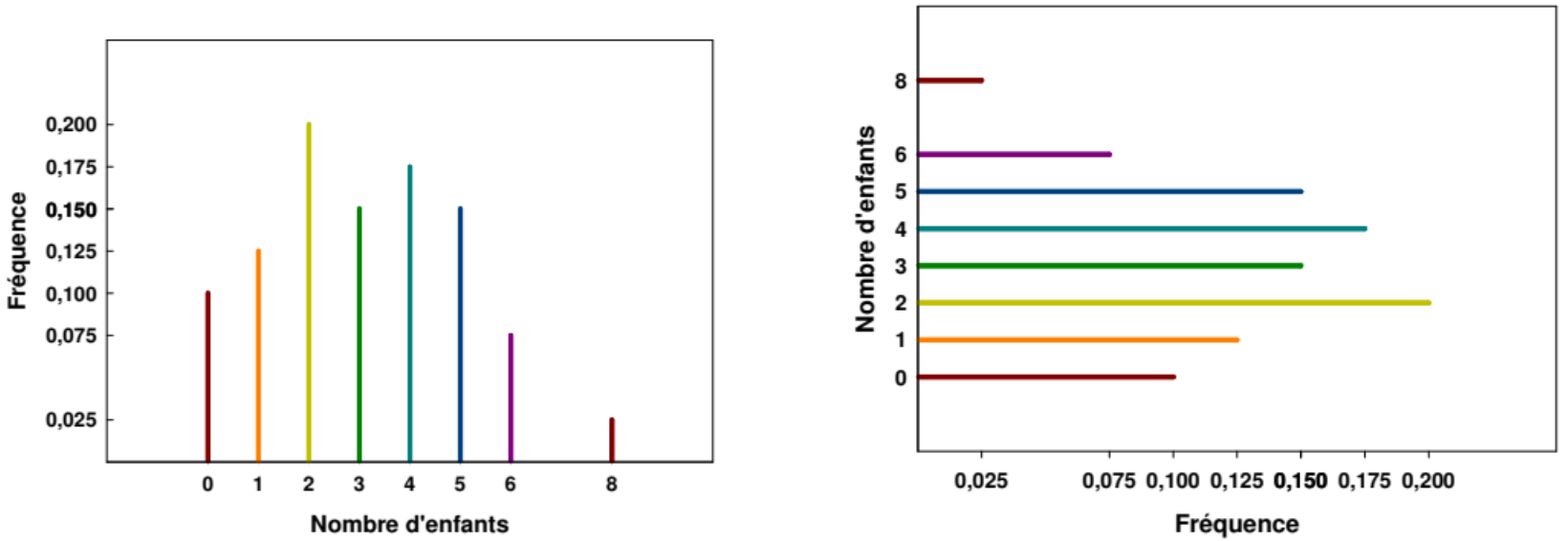


Figure 8 : Diagramme en bâtons vertical et horizontal des fréquences

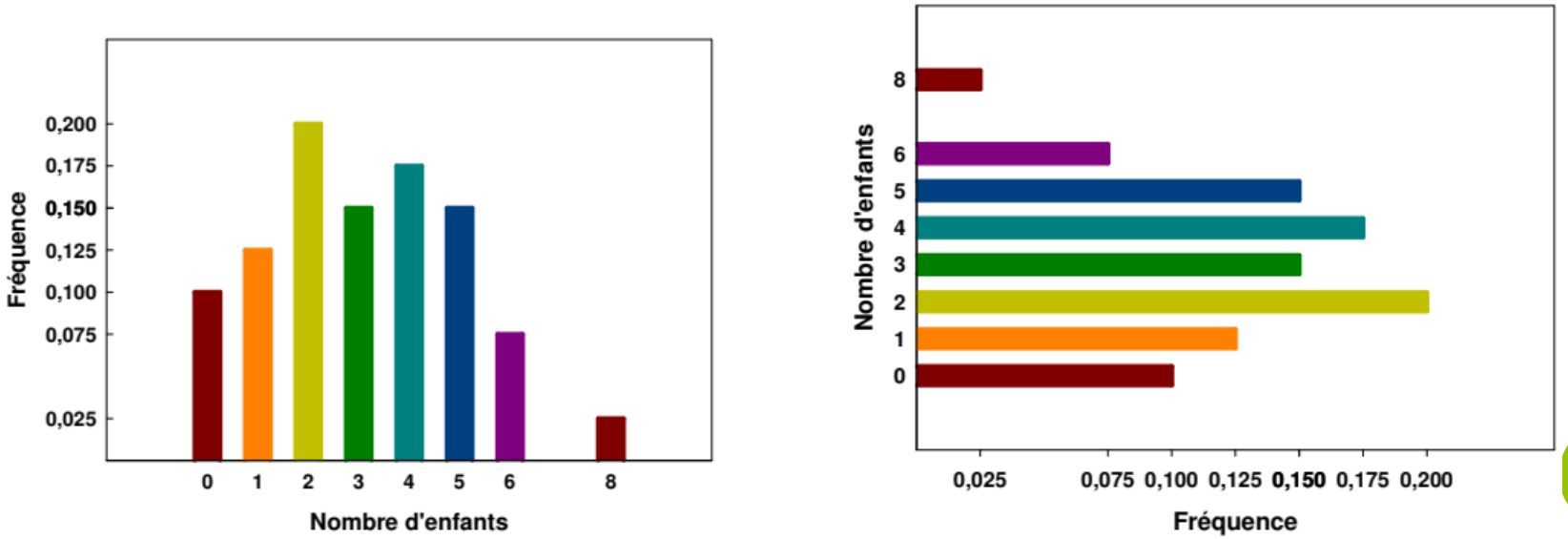
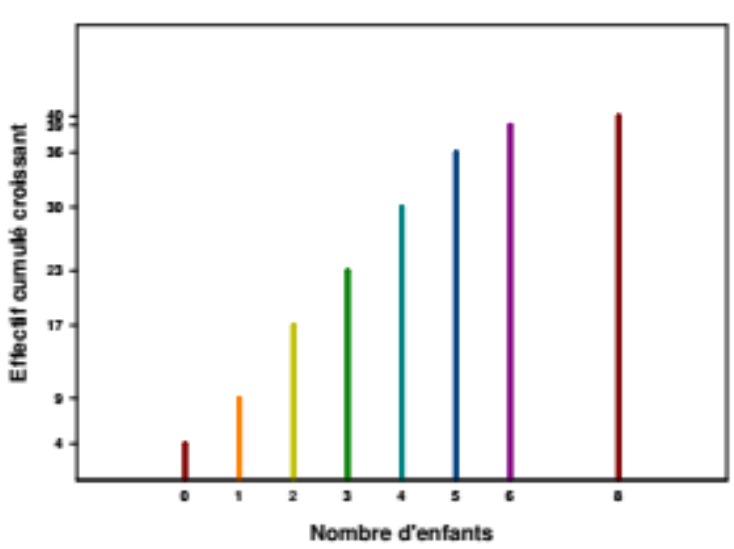


Figure 9 : Diagramme en barres vertical et horizontal des fréquences



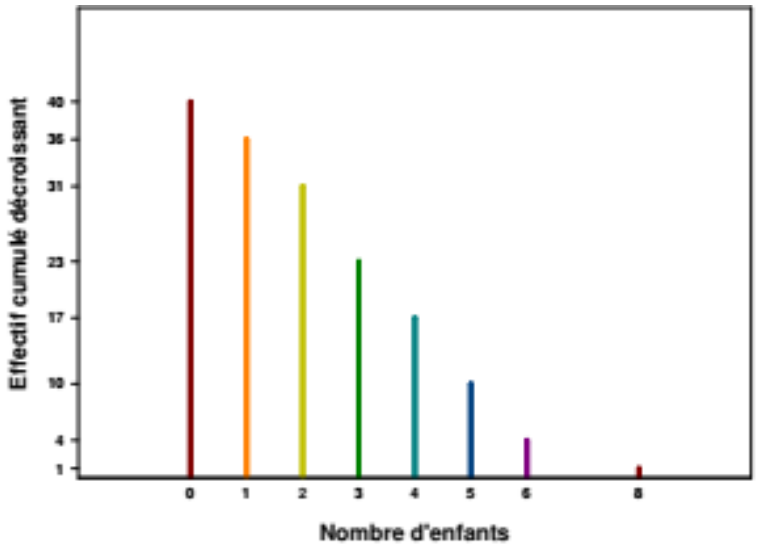
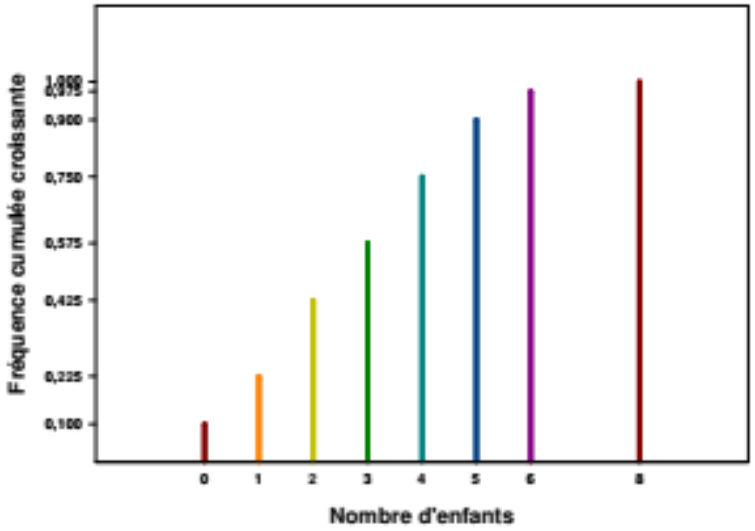


Figure 10 : Diagramme en bâtons vertical des effectifs cumulés croissants et décroissants



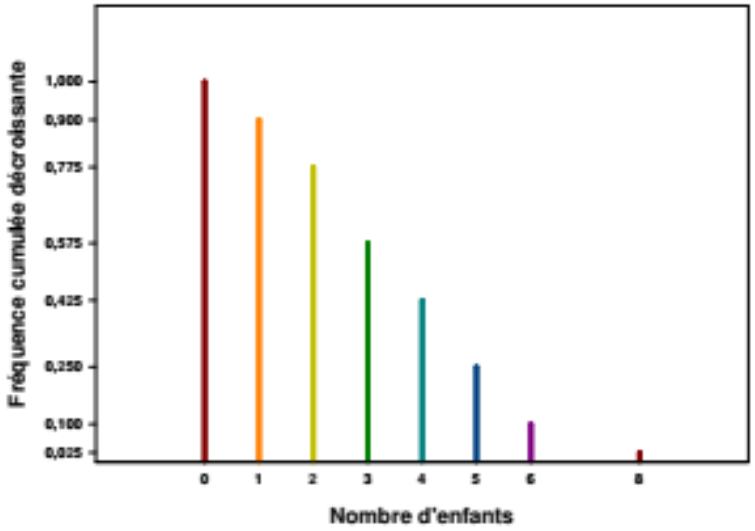


Figure 11 : Diagramme en bâtons vertical des fréquences cumulées croissantes et décroissantes

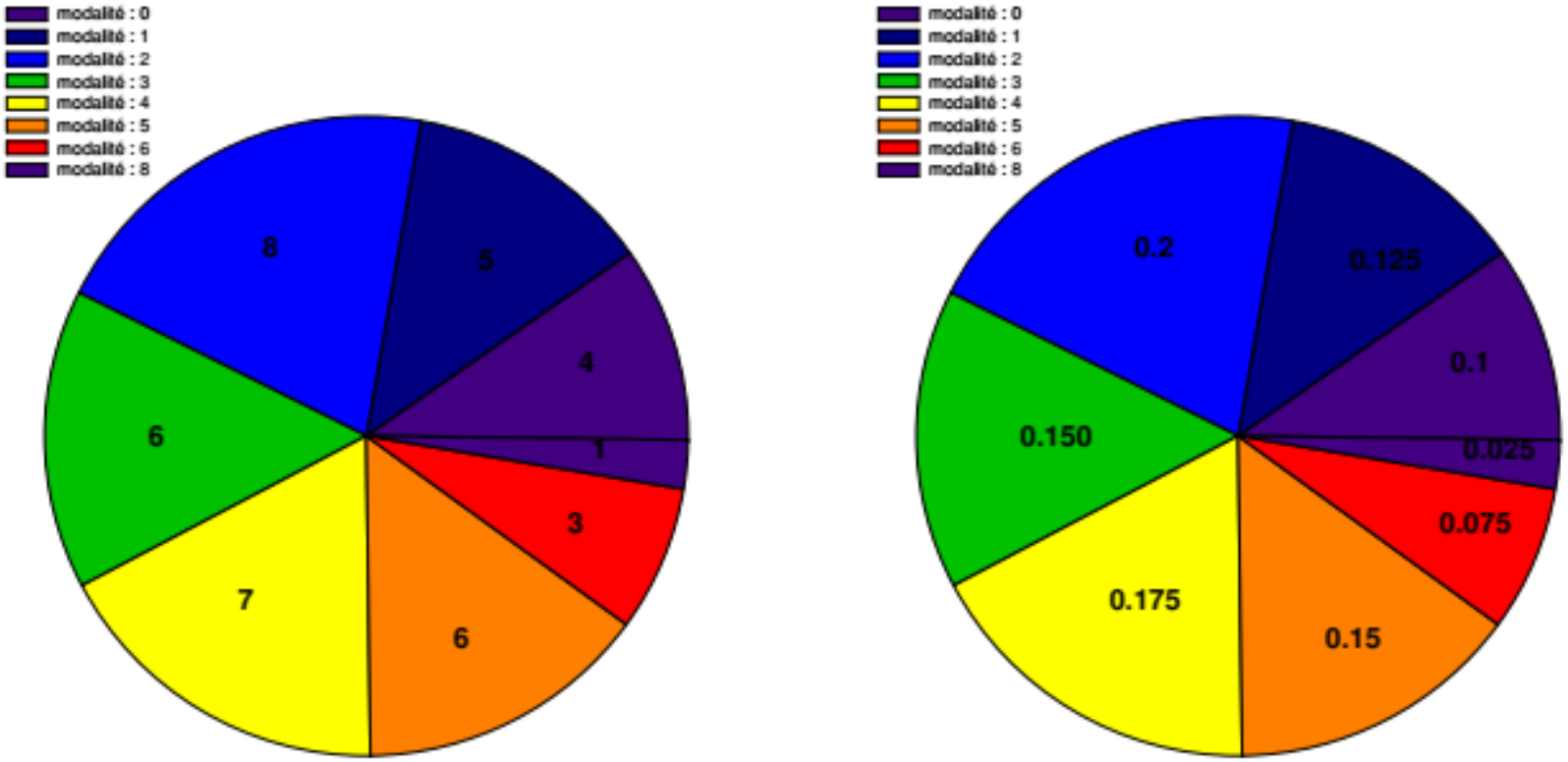


Figure 12 : Diagrammes circulaires des effectifs et des fréquences

I.4.3 Variable statistique quantitative continue

Lorsque la variable est continue, on la représente par un graphique qui exprime cette continuité:

- L'histogramme des effectifs et des fréquences.
- L'histogramme des densités d'effectifs et de fréquences qui nous servira plutard pour le calcul du mode.
- Le polygone des effectifs ou des fréquences et la courbe des fréquences cumulées qui nous sera utile lors du calcul de la médiane ainsi que les quartiles.
- Le diagramme cumulatif des effectifs ou des fréquences.

➢ Histogramme

- Il permet de représenter graphiquement les variables quantitatives continues, sous forme de classes de valeurs. A la différence des variables aléatoires discrètes, une classe donnée ne contient pas une seule valeur mais une infinité de valeurs possibles sur un intervalle défini appelé ***intervalle de classe***.
- Cet intervalle permet de définir également une ***amplitude de classe*** (différence entre les valeurs supérieure et inférieure de la classe). La valeur centrale de la classe est appelée ***centre de classe***.
- Un histogramme consiste en une série de rectangles jointifs et il comprend :
  - ***en abscisses***, les bornes de chaque intervalle ou la base de chaque rectangle correspondant à l'amplitude de la classe
  - ***en ordonnées***, les effectifs ou les fréquences proportionnels à la hauteur du rectangle. La surface de chacun des rectangles, si l'amplitude de classe est constante est alors proportionnelle à l'effectif de la classe.

- On distingue deux cas selon que les amplitudes des classes sont constantes ou variables.
  - ***Amplitudes constantes*** **:** Dans ce cas, les hauteurs des rectangles sont proportionnelles aux effectifs ou aux fréquences.
  - ***Amplitudes variables*** : La hauteur des rectangles n'est plus dans ce cas proportionnelle aux effectifs ou fréquences mais aux effectifs ou fréquences corrigés. Il s'agit en premier lieu de calculer les amplitudes des classes, puis de diviser les effectifs ou les fréquences par les amplitudes de classes respectives.

- On parle alors d'un ***histogramme des densités d'effectifs ou de fréquences***. La densité d'effectif ou de fréquence correspond respectivement au $\frac{n_i}{a_i}$ ou $\frac{n_i}{a_i\, N}$ où a_i est l'amplitude de classe ou la base du rectangle. La surface de chaque rectangle dans cette représentation graphique est alors égale à la fréquence relative de la classe correspondante et la surface totale des rectangle est égale à 1.

Exemple 8 : Amplitude de classe

La répartition par âge et par sexe de la population féminine en France en milliers au 1er janvier 1981était la suivante :

Amplitude de classe

Effectif corrigé

Tranches d'âge		Effecti n_i	$(n_i)/(a_i)$	réquence f_i	Fréquence corrigée $(n_i)/(a_i\ N)$
[15; 18[	3	37	12.333	0.029	0.009
[18; 21[	3	155	51.666	0.124	0.041
[21; 25[	4	267	66.750	0.214	0.053
[25; 30[	5	241	48.200	0.193	0.038
[30; 40[	10	317	31.700	0.254	0.025
[40; 50[	10	138	13.800	0.110	0.011
[50; 60[	10	74	7.400	0.059	0.005
[60; 70[	10	17	1.700	0.013	0.001
Total		N=1246		1	

➢ Polygone des effectifs ou des fréquences

Nous traçons le polygone à partir de l'histogramme en rejoignant les milieux des arrêts supérieurs des rectangles et nous complétons le polygone aux extrémités avec deux points ayant pour ordonnées zéro et pour abscisses le milieu des classes extrêmes.

Exemple 6 (suite) : *Représentation graphique*

Reprenons l'exemple de la variable Z portant sur la taille en centimètres de 50 étudiants, nous obtenons les histogrammes des effectifs et des fréquences et les polygones correspondants suivants :

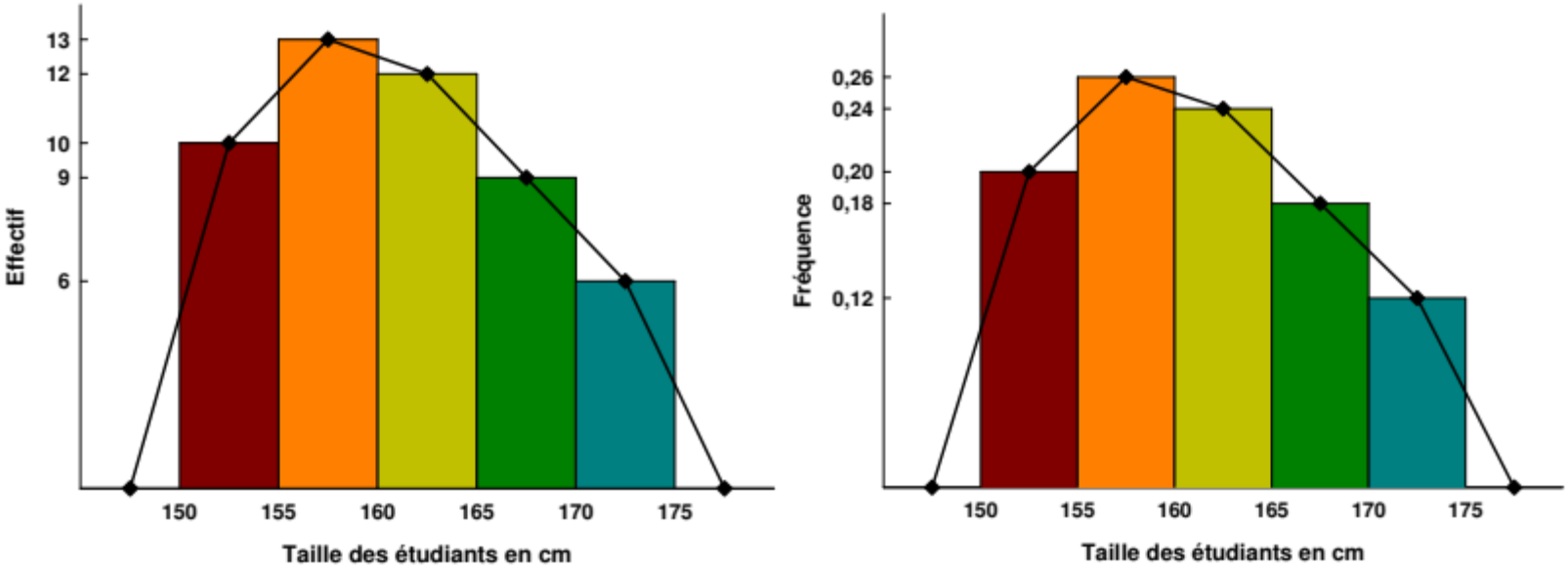


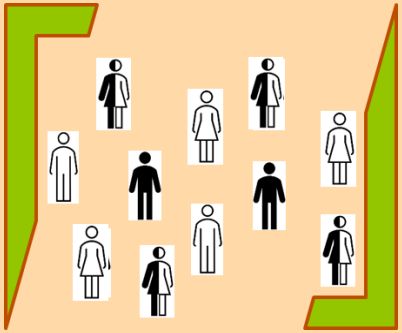
Figure 13 : Histogramme et polygone des effectifs et des fréquences

I.4.3.1 Détermination des classes

- Pour définir les classes d'une série statistique, il faut d'abord prendre en considération
  - Le choix de l'intervalle, nous pouvons distinguer trois catégories : l'intervalle ouvert, fermé et semi-ouvert
  - La partition de ces classes.
- Quant à la partition en classe, pour être efficace, elle doit obéir à certaines règles qu'on essaie de respecter dans la mesure du possible :
  - Le nombre de classe, qu'on note C, ne doit être ni très élevé, ni trop réduit
  - Il doit être supérieur ou égal à cinq
  - Le nombre d'observations par classe doit être, dans la mesure du possible, supérieur ou égal à cinq.
  - Au total, le nombre de classes doit être compris entre 5 et 15 et l'effectif par classe dépassant 5 unités mais en réalité on se trouve souvent devant de petits échantillons et on est obligé de sacrifier l'une des deux conditions.

- Il existe plusieurs méthodes de partition dites ***méthodes statistiques à intervalle régulier***. Cette méthode privilégie la régularité de l'intervalle des classes au dépens des autres règles et essaie de fixer un nombre de classes en fonction de l'effectif global. Nous citons :

- ***Méthode de la racine carrée*** où le nombre de classes est égal à la racine carrée de l'effectif, i.e, $\overline{C} = \sqrt{N}$
- ***Méthode de Brooks-Carruthers*** où le nombre de classes doit être inférieur à 5 fois le logarithme de l'effectif global, i.e, $\overline{C} < 5\ log(N)$
- ***Méthode de Yoles*** où le nombre de classes est fonction de la racine quatrième de l'effectif, i.e, $\overline{C} = 2.5\ N^{1/4}$
- ***Méthode de Huntesberger*** où le nombre de classe est fonction du logarithme de l'effectif, i.e, $\overline{C} = 1 + 3.3\ log(N)$. Cette méthode est présentée parfois sous le nom de ***méthode de Sturge*** avec une formule légèrement différente avec $\overline{C} = 1 + \frac{10\ logN}{3}$
- Il existe bien d'autres méthodes, dont nous nous contentons de les citer, telles que ***la méthode des seuils*** ou la ***méthode synthétique***.



CHAPITRE III

INDICATEURS STATISTIQUES DE POSITION-VALEURS CENTRALES

I.1 Mode

- Le ***mode*** désigne la valeur la plus fréquente d'une variable quelconque de la population.
- Le calcul du mode d'une distribution statistique dépend de la nature continue ou discrète de la variable. Il est parfaitement défini pour une variable qualitative ou une variable quantitative. On le note généralement ***M_o***.

Mode = la valeur dominante d'un ensemble d'observations

- Le mode étant une modalité de la variable, il s'exprime alors dans la même unité que la variable.

I.1.1 Variable statistique qualitative

C'est la valeur de la variable ayant l'effectif ou la fréquence maximal.

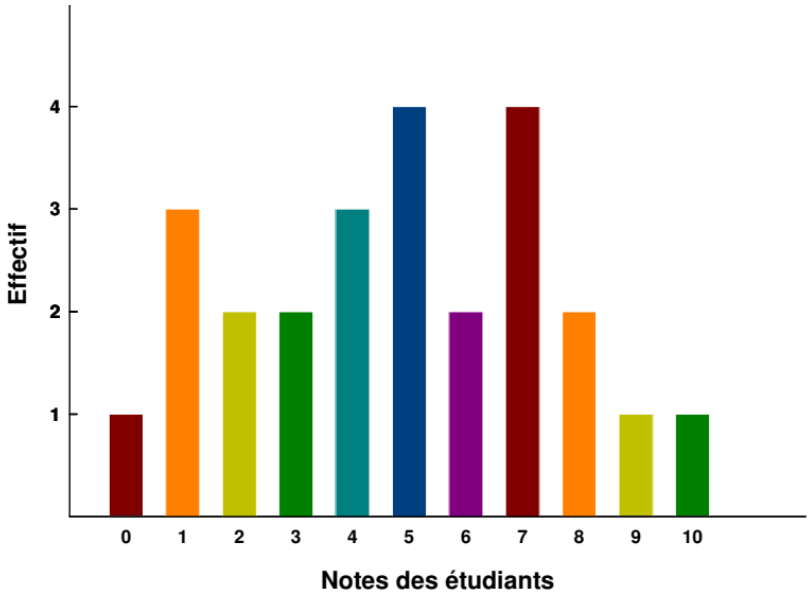
I.1.2 Variable statistique quantitative discrète

Dans le cas d'une série statistique discrète, le mode est la modalité la plus représentée de la variable, autrement dit la modalité ayant l'effectif ou la fréquence les plus grands.

Exemple 9 : Calcul du mode

Soit la distribution statistique d'une population de 25 étudiants et on s'intéresse aux notes attribuées à un examen. Voici le tableau statistique obtenu en classant ces notes par ordre croissant et le diagramme en bâtons des effectifs associé :

Note	Effectif n_i	Fréquence f_i	Fréquence f_i (en %)
0	1	0.04	4
1	3	0.12	12
2	2	0.08	8
3	2	0.08	8
4	3	0.12	12
5	4	0.16	16
6	2	0.08	8
7	4	0.16	16
8	2	0.08	8
9	1	0.04	4
10	1	0.04	4
Total	25	1	100



5 et 7 sont bien les deux modes de cette série car ce sont les deux modalités ayant l'effectif ou la fréquence maximal de cette distribution et sur le diagramme en bâtons, ce sont les 2 modalités associés au bâton le plus élevé. On dit que cette série est ***bimodale*** ou ***plurimodale***.

Exercice 1

Définir le type qualitatif ou quantitatif des caractères suivants :

1) Nom du département de naissance.
2) Sexe.
3) Poids des individus.
4) Prix de l'essence.
5) Couleur des yeux.
6) Nombre d'enfants des couples.
7) Taux de change.
8) Langues parlées officielles d'un pays.
9) Age des enfants des couples.
10) Situation matrimoniale.
11) Salaire des employés.
12) Note de l'épreuve finale en Biostatistique.
13) Moyenne du module Biostatistique.
14) Marque de voiture.
15) Les couleurs par ordre alphabétique.

Caractère qualitatif nominal	Caractère qualitatif ordinal	Caractère quantitatif discret	Caractère quantitatif continu
Nom du département de naissance	Langues parlées officielles d'un pays	Prix de l'essence	Poids des individus
Sexe	Les couleurs par ordre alphabétique	Taux de change	Salaire des employés
Couleur des yeux		Age des enfants des couples	Moyenne du module Biostatistique
Nombre d'enfants des couples		Note de l'épreuve finale en Biostatistique	
Situation matrimoniale			
Marque de voiture			

Exercice 2

Le nombre de frères et sœurs sur un échantillon de 20 étudiants est donné par la liste suivante : 4, 3, 4, 4, 3, 1, 6, 0, 2, 1, 2, 2, 3, 4, 2, 4, 2, 3, 6, 0.

1) Déterminer la population, le caractère et son type ainsi que les modalités et le domaine de cette variable statistique.
2) Classer ces données dans un tableau statistique et déterminer les effectifs, fréquences, effectifs et fréquences cumulés croissants et décroissants.
3) Faire une représentation graphique.
4) Interpréter la ligne 4.

1 Déterminer la population, l'échantillon, le caractère, type de caractère et les modalités

- Population : L'ensemble des 20 étudiants.
- Caractère : Nombre de frères et sœurs par étudiant.
- Type de caractère : Caractère quantitatif discret.
- Modalité : 0, 1, 2, 3, 4, 6.

2 Classer ces données dans un tableau statistique et déterminer les fréquences, effectifs et fréquences cumulés croissants

Nombre de frères et sœurs x_i	Nombre d'étudiants n_i	Effectif cumulé ↗ f_i	Fréquence n_i ↗	Fréquence cumulée ↗ f_i ↗
0	2	2	0.1	0.1
1	2	4	0.1	0.2
2	5	9	0.25	0.45
3	4	13	0.2	0.65
4	5	18	0.25	0.9
6	2	20	0.1	1
Total	20		1	

3 Faire une représentation graphique

Comme il s'agit d'une variable quantitative discrète, on choisit de faire un diagramme en bâton ou un diagramme circulaire :

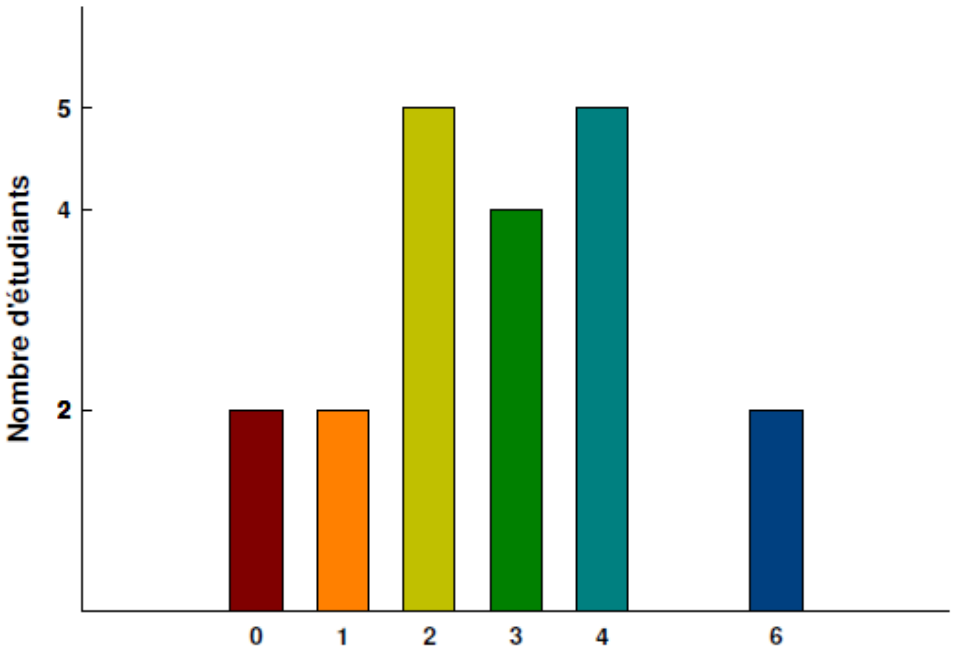


Diagramme en bâtons des effectifs

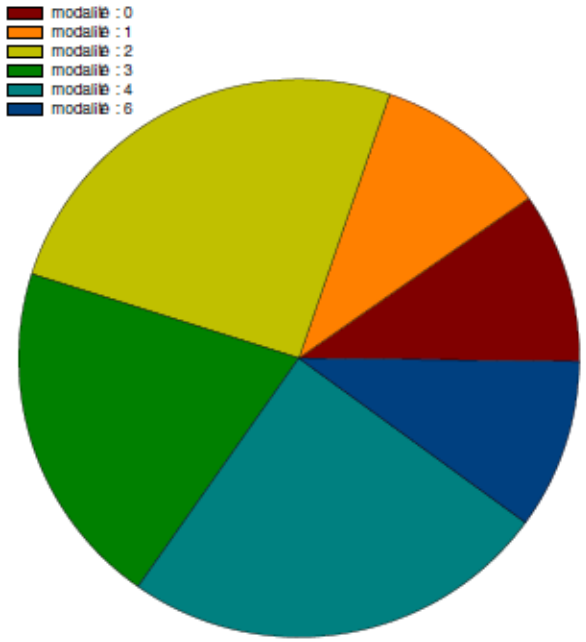


Diagramme circulaire des fréquences

4 Interpréter la "ligne 4"

Nombre de frères et sœurs x_i	Nombre d'étudiants n_i	Effectif cumulé ↗ f_i	Fréquence n_i ↗	Fréquence cumulée ↗ f_i ↗
0	2	2	0.1	0.1
1	2	4	0.1	0.2
2	5	9	0.25	0.45
3	4	13	0.2	0.65
4	5	18	0.25	0.9
6	2	20	0.1	1
Total	20		1	

- 3 étudiants (ce qui correspond à 0.2% de la population étudiée) ont 3 frères ou soeurs et
- 13 étudiants (ce qui correspond à 0.65% de la population étudiée) ont au plus 3 frères ou soeurs et

Exercice 3

Une enquête effectuée auprès des familles selon la date écoulée entre la date de leur mariage et la date d'acquisition de leur premier foyer, se répartissent de la manière suivante :

Date écoulée x_i	Nombre de familles n_i	Fréquence f_i	Effectif cumulé $\nearrow$ $n_i \nearrow$	Fréquence cumulée $\nearrow$ $f_i \nearrow$	Fréquence cumulé $\searrow$ $n_i \searrow$	Fréquence cumulée $\searrow$ $f_i \searrow$
0	92	0.184	92	0.184	500	1
1	73	0.146	165	0.330	408	0.816
2	62	0.124	227	0.454	335	0.670
3	54	0.108	281	0.562	273	0.546
4	46	0.092	327	0.654	219	0.438
5	42	0.084	369	0.738	173	0.346
6	36	0.072	405	0.810	131	0.262
7	30	0.060	435	0.870	95	0.190
8	25					
9	20					
10	9					
11	7					
12	4					
Total	500	1				

Ce tableau doit être compris de la manière suivante : 92 familles ont acquis leur premier foyer pendant la première année de mariage, 73 familles ont acquis leur premier foyer dans leur deuxième année de mariage, 62 familles ont acquis leur premier foyer dans leur troisième année de mariage, etc...

1) Compléter le tableau de l'annexe.
2) Interpréter la "ligne 10".
3) Calculer la moyenne et l'écart-type pour l'ensemble des familles.
4) Calculer la médiane, le mode et l'écart absolu moyen par rapport à la médiane.

1 Compléter le tableau de l'annexe

Date écoulée x_i	Nombre de familles n_i	Fréquence f_i	Effectif cumulé ↗ n_i ↗	Fréquence cumulée ↗ f_i ↗	Fréquence cumulé ↘ n_i ↘	Fréquence cumulée ↘ f_i ↘
0	92	0.184	92	0.184	500	1
1	73	0.146	165	0.330	408	0.816
2	62	0.124	227	0.454	335	0.670
3	54	0.108	281	0.562	273	0.546
4	46	0.092	327	0.654	219	0.438
5	42	0.084	369	0.738	173	0.346
6	36	0.072	405	0.810	131	0.262
7	30	0.060	435	0.870	95	0.190
8	25	0.050	460	0.920	65	0.130
9	20	0.040	480	0.960	40	0.080
10	9	0.018	489	0.978	20	0.040
11	7	0.014	496	0.992	11	0.022
12	4	0.008	500	1	4	0.008
Total	500	1				

2 Interpréter la "ligne 10"

- 20 couples (ce qui correspond à 4% de la population étudiée) ont acquis leur premier foyer pendant leur dixième année de mariage (ou après 9 ans de mariage) et

- 480 couples (ce qui correspond à 96% de la population étudiée) ont acquis leur premier foyer pendant les dix premières années de mariage et
- 40 couples (ce qui correspond à 8% de la population étudiée) ont acquis leur premier foyer à partir de leur dixième année de mariage.

3 Calculer la moyenne et l'écart-type pour l'ensemble des familles

Date écoulée x_i	Nombre de familles n_i	$n_i \cdot x_i$	$x_i - \overline{X}$	$(x_i - \overline{X})^2$	$n_i \cdot (x_i - \overline{X})^2$
0	92	0	-3.534	12.489	1 149.002
1	73	73	-2.534	6.421	468.744
2	62	124	-1.534	2.353	145.895
3	54	162	-0.534	0.285	15.398
4	46	184	0.466	0.217	9.989
5	42	210	1.466	2.149	90.264
6	36	216	2.466	6.081	218.921
7	30	210	3.466	12.013	360.394
8	25	200	4.466	19.945	498.628
9	20	180	5.466	29.877	597.543
10	9	90	6.466	41.809	376.282
11	7	70	7.466	55.741	390.188
12	4	48	8.466	71.673	286.692
Total	500	1 767			4 607.94

La moyenne est donnée par la formule suivante :

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{13} n_i\, x_i = \frac{1767}{500} = 3.534.$$

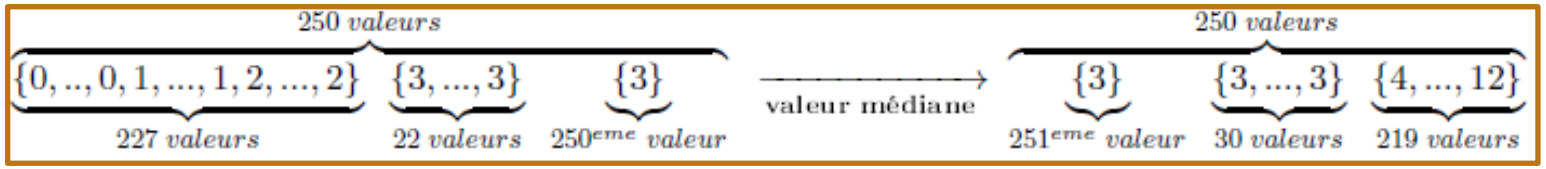
Par conséquent, les couples acquièrent en moyenne leur premier foyer après 3.534 années de mariage.

L'écart-type est donné par la formule suivante :

$$\sigma(X) = \sqrt{V(X)} = \sqrt{\frac{1}{N}\sum_{i}^{13} n_i\,(x_i - \overline{X})^2} = \sqrt{\frac{4607.94}{500}} = 3.035\ annees.$$

4 Calculer la médiane, le mode et l'écart absolu moyen par rapport à la médiane

On a que N=500 et vu que la valeur N/2=250 ne correspond pas exactement à une modalité de la variable et tombent entre deux valeurs de la distribution statistique : 2 et 3, par convention, on retiendra comme valeur médiane la valeur de la variable immédiatement supérieure, d'où Me=3. Voici la répartition des années pour comprendre



Dans le tableau, il n'y a pas de valeur partageant la série statistique en 2 groupes de même effectif. Dans ce cas, l'intervalle médian est [3,3] et on prend pour médiane le centre de cet intervalle Me=3.

Le mode est Mo=0 car c'est la modalité la plus représentée à laquelle on associe l'effectif maximal n_1=92.

L'écart absolu moyen par rapport à la médiane est donnée par la formule suivante :

$$e_X = \frac{1}{N} \sum_i n_i \, |x_i - Me|$$

Date écoulée x_i	Nombre de familles n_i	$x_i - Me$	$n_i \, . \, \lvert x_i - Me \rvert$
0	92	-3	276
1	73	-2	146
2	62	-1	62
3	54	0	0
4	46	1	46
5	42	2	84
6	36	3	108
7	30	4	120
8	25	5	125
9	20	6	120
10	9	7	63
11	7	8	56
12	4	9	36
Total	500		1242

D'où

$$e_X = \frac{1242}{500} = 2.484 \; annees$$

Exercice 4

La répartition des salaires de 500 salariés d'une entreprise se présente ainsi :

Salaires	Salariés
$[1000\,;\,1200[$	25
$[1200\,;\,1500[$	125
$[1500\,;\,1700[$	75
$[1700\,;\,1900[$	150
$[1900\,;\,2200[$	50
$[2200\,;\,2500[$	75

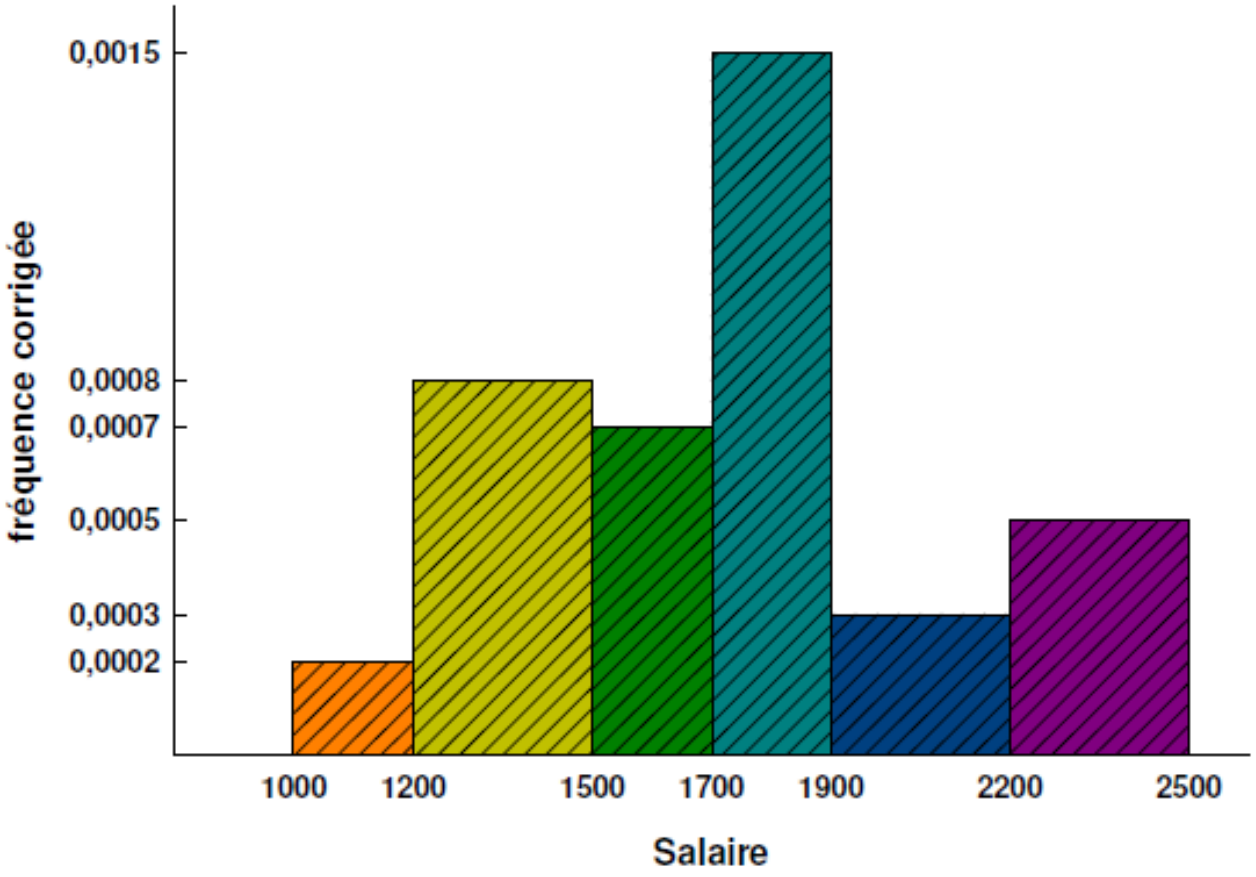
1) Tracer l'histogramme de cette distribution.
2) Déterminer graphiquement et analytiquement le mode.
3) Tracer la courbe cumulative.
4) Déterminer graphiquement et analytiquement la médiane et interpréter le résultat.
5) Calculer la moyenne, la variance et l'écart-type de cette distribution.
6) Déterminer graphiquement et analytiquement le $4^{ème}$ décile, le $3^{ème}$ quartile et le 45éme centile et interpréter.

Tracer l'histogramme de cette distribution

Dressons le tableau statistique attribué à cette distribution. Vu que les amplitudes des classes sont variables, il faut au préalable corriger les effectifs et les fréquences afin de tracer l'histogramme des fréquences.

Salaires	Amplitude de classe a_i	Nombre de salariés n_i	Effectif corrigé $(n_i)/(a_i)$	Fréquence f_i	Fréquence corrigée $(f_i)/(a_i)$
$[1000\,;\,1200[$	200	25	0.125	0.05	0.0002
$[1200\,;\,1500[$	300	125	0.416	0.25	0.0008
$[1500\,;\,1700[$	200	75	0.375	0.15	0.0007
$[1700\,;\,1900[$	200	150	0.750	0.30	0.0015
$[1900\,;\,2200[$	300	50	0.166	0.10	0.0003
$[2200\,;\,2500[$	300	75	0.250	0.15	0.0005
Total		500		1	

Nous obtenons l'histogramme des fréquences corrigées suivant :



2 Déterminer graphiquement et analytiquement le mode

La classe modale, à laquelle est associé l'effectif ou la fréquence corrigés les plus importants est la classe [1700;1900] euros.

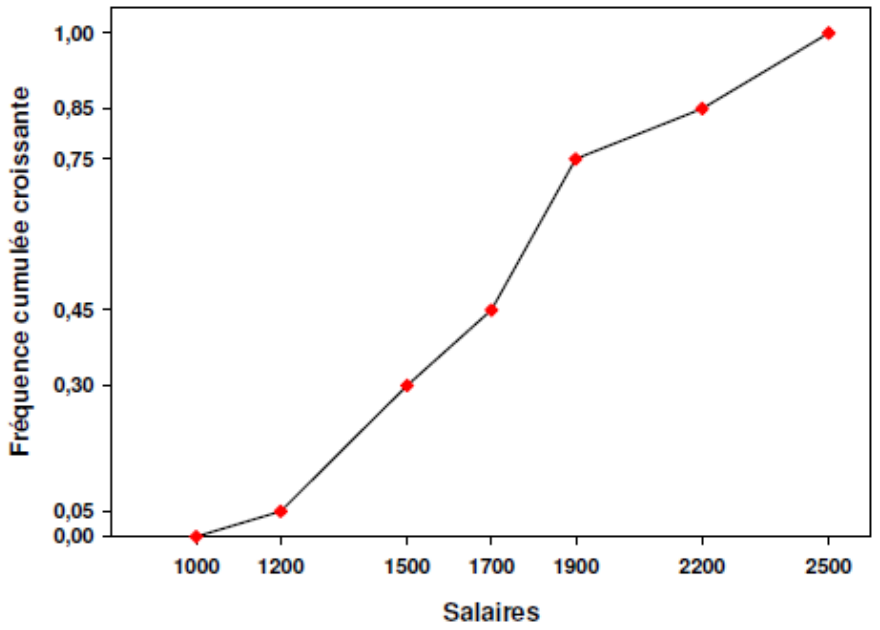
Graphiquement, la première méthode pour le calcul du mode consiste à joindre les segments [AB] et [CD] et déterminer le point d'intersection de ces deux segments. L'abscisse de ce point sera le mode Mo=1800 euros. Une autre méthode graphique de calcul du mode consiste à prendre le centre de la classe modale Mo=1800 euros.

Analytiquement, le mode est défini comme suit:

$$M_0 = 1700 + \frac{(0.0015 - 0.0007)}{(0.0015 - 0.0007) + (0.0015 - 0.0003)}(1900 - 1700) = 1778.125\ euros.$$

3 Tracer la courbe cumulative

Courbe cumulative des fréquences de la variable statistique "salaires"



4 Déterminer graphiquement et analytiquement la médiane et interpréter le résultat

Graphiquement sur la courbe cumulative, la médiane est l'abscisse du point dont l'ordonnée est **0.5**, d'où $Me = 1740$ euros. Analytiquement d'après la colonne des fréquences cumulées,

Salaires	**Fréquence** f_i	**Fréquence cumulée** $\nearrow$ $f_i \nearrow$
$]1000\,;\,1200]$	0.05	0.05
$]1200\,;\,1500]$	0.25	0.30
$]1500\,;\,1700]$	0.15	0.45
$]1700\,;\,1900]$	0.30	0.75
$]1900\,;\,2200]$	0.10	0.85
$]2200\,;\,2500]$	0.15	1
Total	1	

ane est $[1700; 190]$ euros et la médiane est définie comme suit :

$$\frac{Me - 1700}{1900 - 1700} = \frac{0.50 - 0.45}{0.75 - 0.45}.$$

733.33 euros. On a donc parmi les 500 salariés, autant de salariés dont le salaire est **33.33** euros, que de salariés dont le salaire est supérieur à cette valeur médiane.

5 Calculer la moyenne, la variance et l'écart-type de cette distribution

Salaires	**Centre de classe** c_i	**Nombre de salariés** n_i	$n_i \,.\, c_i$	$c_i - \overline{X}$	$(c_i - \overline{X})^2$	$n_i \,.\, (c_i - \overline{X})^2$
$[1000\,;\,1200[$	1100	25	27 500	- 630	396 900	9 922 500
$[1200\,;\,1500[$	1350	125	168 750	- 380	144 400	18 050 000
$[1500\,;\,1700[$	1600	75	120 000	- 130	16 900	1 267 500
$[1700\,;\,1900[$	1800	150	270 000	70	4 900	735 000
$[1900\,;\,2200[$	2050	50	102 500	320	102 400	5 120 000
$[2200\,;\,2500[$	2350	75	176 250	620	384 400	28 830 000
Total		500	865 000			63 925 000

La moyenne est donnée par la formule suivante :

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{6} n_i\, c_i = \frac{865000}{500} = 1730.$$

Par conséquent, le salaire moyen des salariés de cet entreprise est égal à 1730 euros.

L'écart-type est donné par la formule suivante :

$$\sigma(X) = \sqrt{V(X)} = \sqrt{\frac{1}{N}\sum\nolimits_i^6 n_i\,(c_i - \overline{X})^2} = \sqrt{\frac{63925000}{500}} = 357.561\ \textit{euros}.$$

6 Déterminer graphiquement et analytiquement le 4^{eme} décile, le 3^{eme} quartile et le 45^{eme} centile et interpréter

Le 4^{eme} décile est la valeur de la variable telle que 40% des observations lui sont inférieures. Graphiquement sur la courbe cumulative, le 4^{eme} décile est l'abscisse du point dont l'ordonnée est 0.4, d'où $D_4 = 1640$ euros. Analytiquement d'après la colonne des fréquences cumulées, la classe est $[1500; 1700]$ euros et il est défini comme suit :

$$\frac{D_4 - 1500}{1700 - 1500} = \frac{0.40 - 0.30}{0.45 - 0.30}.$$

Soit $D_4 = 1633.33$ euros. On a donc parmi les 500 salariés, 40% des salariés ont un salaire inférieur à 1633.33 euros, et 60% des salariés ont un salaire supérieur à cette valeur.

Le 3^{eme} quartile est la valeur de la variable telle que 75% des observations lui sont inférieures. Graphiquement sur la courbe cumulative, le 3^{eme} quartile est l'abscisse du point dont l'ordonnée est 0.75, d'où $Q_3 = 1900$ euros. Analytiquement d'après la colonne des fréquences cumulées, la classe est $[1700; 1900]$ euros et il est défini comme suit :

$$\frac{Q_3 - 1700}{1900 - 1700} = \frac{0.75 - 0.45}{0.75 - 0.45}.$$

Soit $Q_3 = 1900$ euros. On a donc parmi les 500 salariés, 75% des salariés ont un salaire inférieur à 1900 euros, et 25% des salariés ont un salaire supérieur à cette valeur.

Le 45^{eme} centile est la valeur de la variable telle que 45% des observations lui sont inférieures. Graphiquement sur la courbe cumulative, le 45^{eme} centile est l'abscisse du point dont l'ordonnée est 0.45, d'où $C_{45} = 1700$ euros. Analytiquement d'après la colonne des fréquences cumulées, la classe est $[1500; 1700]$ euros et il est défini comme suit :

$$\frac{C_{45} - 1500}{1700 - 1500} = \frac{0.45 - 0.30}{0.45 - 0.30}.$$

Soit $C_{45} = 1700$ euros euros. On a donc parmi les 500 salariés, 45% des salariés ont un salaire inférieur à 1700 euros, et 55% des salariés ont un salaire supérieur à cette valeur.

Exercice 5

On a relevé sur 20 auto-écoles de Marseille, le montant proposé pour la formation complète au permis de conduire «B». Les résultats nets obtenus en € sont les suivants : 1240 ; 790 ; 900 ; 770 ; 1070 ; 835 ; 720 ; 850 ; 950 ; 1200, 1340 ; 990 ; 800 ; 790 ; 1070 ; 850 ; 750 ; 850 ; 950 ; 1150.

Voici une répartition suivant quatre classes :

Montant en €	Effectif
$[0 ; 800[$	5
$[800 ; 900[$	5
$[900 ; 1050[$	4
$[1050 ; 1350[$	6

1. Quels sont la population et les individus étudiés ? Donner la taille de la population.
2. Quelle est la variable statistique étudiée ? Préciser sa nature.
3. Pour cette variable, quelle est la série statistique et quelle est la distribution d'effectifs.
4. Pour la distribution, quelles sont les modalités de la variable statistique ? Combien sont-elles ?
5. Construire l'histogramme de cette distribution en indiquant les axes des abscisses et des ordonnées ainsi que le polygone des effectifs.
6. Construire les courbes cumulatives (fonctions de répartition) des effectifs et des fréquences.
7. Déterminer graphiquement et analytiquement le mode.
8. Déterminer graphiquement et analytiquement la médiane et interpréter le résultat.
9. Calculer la moyenne, la variance et l'écart-type de cette distribution.

1 Quels sont la population et les individus étudiés ? Donner la taille de la population

Population : L'ensemble des auto-écoles de Marseille.
Individus : une auto-école quelconque de Marseille.
Taille de la population : 20.

2 Quelle est la variable statistique étudiée ? Préciser sa nature

la variable statistique étudiée est le montant proposé pour la formation complète au permis de conduire «B». C'est une variable quantitative continue.

3 Pour cette variable, quelle est la série statistique et quelle est la distribution d'effectifs

Soit la fonction $X : P \to \mathbb{R}$, qui à chaque auto-école $w \in P$, $X(w)$ représente le montant proposé par w.

X(w)={1240 ; 790 ; 900 ; 770 ; 1070 ; 835 ; 720 ; 850 ; 950 ; 1200, 1340 ; 990 ; 800 ; 790 ; 1070 ; 850 ; 750 ; 850 ; 950 ; 1150}.

4 Pour la distribution, quelles sont les modalités de la variable statistique ? Combien sont-elles ?

Les modalités de cette variable sont les classes suivantes : $[0; 800[$, $[800; 900[$, $[900; 1050[$ et $[1050; 1350[$.
Elles sont 4 modalités.