

Eléments de statistique descriptive

Ter Tiero Elias DAH MD,PhD
AHU Santé Publique – Option Epidémiologie
Université de Ouahigouya
Desiré Lucien Dahourou
Chargé de Recherche Santé Publique-Option Epiémiologie
IRSS/CNRST

Plan

1. Types de variables
2. Description d'une distribution statistique
 - 2.1. Présentation de données brutes
 - 2.2. Mesures de réduction statistique
 - 2.2. Tableaux
 - 2.3. Graphiques ou figures

I. Types de variables

Types de variables

Deux types de variables en statistique(1)

▪ Qualitative ou catégorielle

- Qualifie un objet ou un évènement
- Les réponses aux questions permettant de recueillir ces variables comportent des catégories ou modalités
- Les réponses ne sont pas numériques. Cependant lors du codage des modalités de réponses, par souci de commodité et par nécessité informatique, l'on peut précéder chacune d'elle par un nombre entier (valeur numérique) différent.

NB: Pour une variable qualitative, l'écart entre 2 valeurs utilisées pour le codage ne quantifie pas l'écart entre les 2 catégories correspondantes.

Ex: Quel est votre sexe? Femme Homme

Quel est l'âge de votre enfant? 0-3 mois 4-11 mois 1-3 ans 4-10 ans

Quel(s) signe(s) aviez-vous? 1.Fièvre 2.Céphalées 3.Sueurs 4.Autres (Préciser)

Types de variables

Deux types de variables en statistique (2)

■ Quantitative ou numérique

- Représente la mesure d'une quantité
- Prend des valeurs numériques qui ont une signification concrète (ex. la taille, le poids...)
- Un calcul mathématique (somme, moyenne) peut être réalisé à partir des réponses d'un groupe de sujets.
- Les questions introduites par l'adverbe "combien" recueille des variables quantitatives.

NB: L'écart entre 2 valeurs d'une variable quantitative est interpretable et peut être comparé avec 2 autres valeurs.

Ex. Pour une taille, l'écart entre 1,60 m et 1,65 m est le même qu'entre 1,63 m et 1,68.

Exemple:

Question	Information recueillie	Variable	Valeurs prises ou modalités
Combien d'enfants avez-vous?	Nb d'enfants/personne	Nombre d'enfant	1, 2, 3,n
Combien pesez-vous?	Poids du sujet	Poids en kg	$20 < k < 125$
Quel est votre PAS	Pression artérielle systolique/ personne	PAS (mm Hg)	

Variables qualitatives

Deux types

▪ Les variables qualitatives ordinales

Les modalités de réponses sont ordonnées selon un ordre logique.

Exemple:

- Niveau de satisfaction vis à vis du cours de biostatistique : très satisfait, satisfait, moyennement satisfait, pas du tout satisfait, sans opinion.
- Le stade de classification OMS de l'infection par le VIH (codé de 1 à 4). Attention, ce codage ne doit pas être confondu avec une réponse numérique.

▪ Les variables qualitatives nominales

Les catégories ne sont pas ordonnées. Elles nomment un objet, un évènement ou une chose.

Exemple:

- La profession des personnes entrant aux urgences médicales du CHUR de Ouahigouya. Les modalités (ou catégories) possibles de réponses peuvent être: agriculteur, fonctionnaire, étudiant, sans emploi fixe...

NB: La **variable qualitative binaire ou dichotomique** n'admet que 2 modalités de réponses, habituellement codées **1** et **0**. Elle est à première vue un cas particulier de variable qualitative nominale. Cependant, elle est peut être considérée d'ordinaire si l'on admet que la modalité codée 1 est plus grande que celle codée 0; voire être considérée de variable quantitative discrète si l'on décide que la distance entre les 2 valeurs possibles est égale à 1.

Ex: Sexe: 0= homme; 1= femme. Infection par le Sars-CoV-2: 1= oui; 2=non;

Variables quantitatives (1)

Deux types

- **Les variables quantitatives discrètes**

Elles ont un nombre fini ou dénombrable (que l'on peut énumérer) de valeurs possibles. Ces valeurs sont distinctes et séparées, aucune valeur intermédiaire n'est possible.

Exemple:

- Le nombre d'enfants dans une famille. Le nombre de patients qui consultent dans le service de cardiologie par jour.

NB: La distinction entre certaines variables quantitatives discrètes et les variables ordinales ne sont pas toujours évidentes (Ex: nombre d'enfants et stade d'un cancer codé 1,2,3,4). Pour les distinguer, l'on peut faire un test simple. Pour une variable catégorielle ordinale, chaque différence entre les catégories ne signifie pas la même chose. En revanche, pour une variable quantitative discrète, chaque différence entre deux catégories successives a toujours la même signification sur toute l'étendue des valeurs.

Ex: Pour la variable « stade de cancer », on ne peut pas dire que le stade 2 est deux fois plus grave que le stade 1 ; c'est donc une variable catégorielle ordinale. Pour la variable « nombre d'enfants », on peut dire que deux enfants, c'est deux fois plus que un, et que trois enfants c'est trois fois plus que un ; c'est donc une variable quantitative discrète.

Variables quantitatives (2)

Deux types

- **Les variables quantitatives continues**

Elles peuvent prendre un nombre infini de valeurs entre deux bornes définies par la valeur minimale et la valeur maximale. Il s'agit de toutes les valeurs qui mesurent des quantités ou grandeurs physiques (poids, taille, dosages biologique...).

Souvent l'information exacte ne peut pas être retranscrite car les instruments de mesure arrondissent l'information à un certain niveau de précision.

Par exemple, si la mesure du poids se fait au moyen d'un pèse-personne, cette information est souvent arrondie aux centaines de grammes près. Finalement, on regroupe les mesures d'individus par paquets de cent grammes. Les sujets qui ont un poids compris entre 65,050 kg et 65,150 kg, par exemple, obtiennent une valeur de 65,100 kg sur le pèse-personne et c'est cette valeur que l'on retient. On dit alors que l'on **discrétise** la mesure. De même une information telle que l'âge est souvent recueillie sous la forme d'une variable discrète avec comme unité l'année (22, 23, 24 ans, etc.).

Il est donc assez évident que toute variable, que l'on peut recueillir sous un format de variable continue, peut aussi être recueillie sous un format de variable discrète, cela entraînant une perte d'information.

Différents types de variables et quelques exemples

Variables qualitatives	Variables quantitatives
<p>Ordinales</p> <ul style="list-style-type: none">• Niveau d'études• Stade de gravité d'une pathologie	<p>Discrètes</p> <ul style="list-style-type: none">• Nombre d'enfants d'une famille• Age en années
<p>Nominales</p> <ul style="list-style-type: none">• Groupe sanguin /Rhésus• Profession	<p>Continues</p> <ul style="list-style-type: none">• Poids• Dosages biologiques
<p>Binaire</p> <ul style="list-style-type: none">• Sexe• Malades/Non malades	

II. Description d'une distribution statistique

II.1. Données brutes

Présentation de données brutes

Il s'agit de présenter l'ensemble ou une partie des variables et des effectifs correspondants. La présentation est le plus souvent pour la personne chargée de l'analyse des données. Elle permet d'avoir une idée globale des données recueillies et leur qualité. Ceci constitue rarement la seule forme de présentation des données statistiques, mais plutôt une première étape.

■ Exemples

– N° patient	Sexe	Age (ans)	Poids (kg)	Diabète
1	F	69	75	O
2	F	75	58	N
3	M	60	88	O
4	F	72	70	O
5	M	69	68	N
6	F	71	63	N
7	M	71	75	N
8	M	73	72	N
9	F	77	60	N
10	F	68	56	N

– Ages des sujets (ans) : 69, 75, 60, 72, 69, 71, 71, 73, 77, 68.

II.2. Réduction ou résumé des données

Mesures (ou statistiques) de description

2 types de paramètres permettent de résumer la distribution d'une variable en fonction des valeurs observées. Ils sont fonction du type de variable.

Paramètres de position ou de tendance centrale

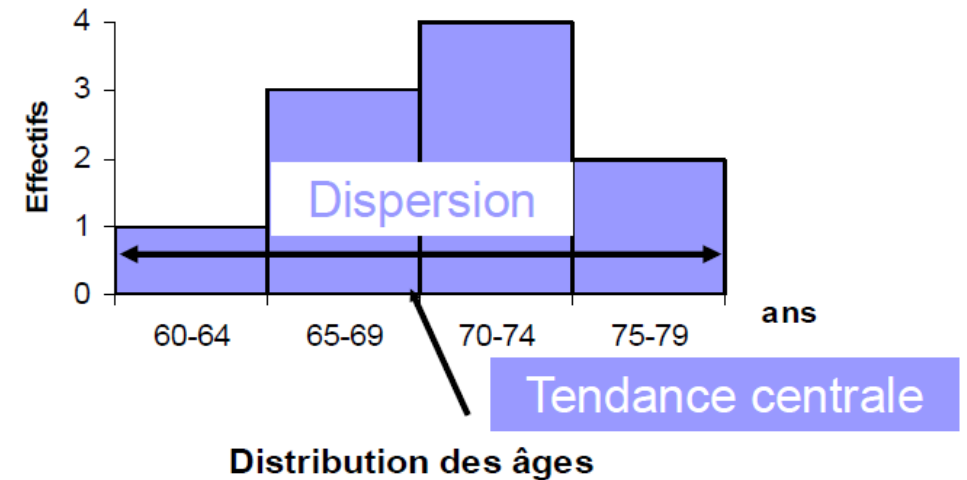
= valeur unique qui résume une série de données

- Pour une variable quantitative
 - Mode
 - Médiane
 - Moyenne
- Pour une variable qualitative
 - Fréquence

Paramètres de dispersion

= valeurs qui informent sur la variation autour de la tendance centrale

- Pour une variable quantitative
 - Etendue
 - Variance/ Ecart-type
 - Percentiles
- Pour une variable qualitative
 - Intervalle de confiance



Tendance centrale

Mode (1)

On appelle **mode** la valeur prise le plus fréquemment par la variable, ayant donc le plus grand effectif (ce terme est inspiré du mot « la mode » caractérisant la valeur la plus « en vogue »).

Si plusieurs valeurs de la variable ont l'effectif maximal, on dit que la distribution est **plurimodale**. On parle de distribution **bimodale** si elle possède deux modes.

Si la variable est continue, on appelle **classe modale** la classe ayant le plus grand effectif. Cependant pour que cela ait un sens, il est nécessaire que toutes les classes soient de même amplitude.

Exemple

Durant une enquête réalisée auprès de 200 familles, on recueille l'information de la superficie de leur logement ainsi que le nombre de voitures possédées par chaque famille.

La variable « nombre de voitures » est une variable quantitative discrète, tandis que la variable « superficie du logement » est quantitative continue. Les résultats suivants sont relevés :

Tendance centrale

Mode (2)

Répartition des 200 familles selon leur nombre de voitures

Nombre de voitures	Effectifs
0	20
1	100
2	67
3	10
Au moins 4	3

Le mode de la variable « nombre de voitures » est la valeur 1 car il s'agit du nombre de voitures le plus fréquent parmi les 200 familles.

Répartition des 200 familles selon la superficie du logement

Superficie de logements (m ²)	Effectifs
[20 ; 50[15
[50 ; 80[50
[80 ; 110[88
[110 ; 140[35
Au moins 140	15

La classe modale est [80 ; 110[

Tendance centrale

Médiane (1)

C'est la valeur de la variable qui divise en 2 parties égales une série ordonnées de valeurs

- Les valeurs observées sont ordonnées
- 50% des valeurs sont inférieures à la médiane et 50% sont supérieures

Méthode de détermination

1) Il faut ordonner les N valeurs de la plus petite à la plus grande, la plus petite valeur aura le rang 1 et la plus grande le rang N.

2) -Si l'effectif total N est un nombre impair on peut écrire que $N=2*k+1$, la valeur de l'observation ayant le rang (k+1) représente la **médiane**.

- Si l'effectif total est un nombre pair on peut écrire que $N=2*k$, la valeur de la médiane est donc comprise entre la valeur de l'observation ayant le rang k et la valeur de l'observation ayant le rang (k+1),

a) soit les valeurs des observations de rang k et de rang (k+1) sont identiques et alors la médiane est égale à cette valeur,

b) soit les deux valeurs sont différentes et alors on indique **l'intervalle médian**, défini par **[valeur de l'observation k ; valeur de l'observation (k+1)[**

Tendance centrale

Médiane (2)

Exemples

1) Les notes (sur 20) à un devoir de français d'une classe de 20 élèves âgés de 10 ans sont les suivantes : 5, 6, 6, 8, 9, 9, 9, 10, 11, 12, 12, 12, 12, 13, 13, 14, 15, 17, 18, 19.

L'effectif total 20, est un nombre pair, la médiane est donc comprise dans l'intervalle formé par la 10ème et la 11ème valeur, c'est à dire ici la valeur 12 et la valeur 12, la médiane est donc égale à 12.

2) Pour définir la médiane de la variable « nombre de voitures » de l'enquête réalisée auprès de 200 familles, on peut utiliser le tableau obtenu précédemment et ajouter une colonne représentant les effectifs cumulés. La valeur représentant la médiane est comprise entre la valeur du 100ème sujet et du 101ème sujet.

Répartition des 200 familles selon leur nombre de voitures et effectifs cumulés

Nombre de voitures	Effectifs	Effectifs cumulés
0	20	20
1	100	120
2	67	187
3	10	197
Au moins 4	3	200

Les sujets classés 21ème à 120ème ont la valeur 1 pour la variable « nombre de voitures », donc la médiane a pour valeur 1 car le 100ème sujet a la valeur 1.

Tendance centrale

Médiane (3)

Exemples

- Nombre d'enfants dans chacune des 10 familles

2, 8, 1, 1, 4, 5, 2, 3, 0, 1

Nombre médian d'enfants / famille ?

Valeurs :	0	1	1	1	2	2	3	4	5	8
Rangs :	1	2	3	4	5	6	7	8	9	10

A red arrow points from the number **2** to the vertical line between the 5th and 6th columns of the table above.

N=10

Valeurs :	0	1	1	1	2	3	3	4	5	8
Rangs :	1	2	3	4	5	6	7	8	9	10

A red arrow points from the number **2,5** to the vertical line between the 5th and 6th columns of the table above.

N=11

Valeurs :	0	1	1	1	2	3	3	3	4	5	8
Rangs :	1	2	3	4	5	6	7	8	9	10	11

A red arrow points from the number **3** to the vertical line between the 6th and 7th columns of the table above.

Tendance centrale

Médiane (4)

Méthode de détermination de la médiane lorsque les N valeurs ne sont connues que par leur appartenance à un intervalle

Si la médiane appartient à un intervalle $[a, b[$ alors on peut conclure que l'intervalle médian est égal à l'intervalle $[a ; b[$

ou estimer la médiane, en supposant une répartition à l'intérieur d'une classe linéaire et en utilisant une méthode d'*interpolation linéaire (formule hors programme)*

Tendance centrale

Moyenne (1)

Soit X une variable aléatoire, on note \bar{X} la moyenne arithmétique obtenue par le calcul suivant :

$$\bar{X} = \frac{\sum_{i=1}^N x_i}{N} \quad \text{moyenne arithmétique simple (N le nombre de sujets).}$$
$$\text{ou } \bar{X} = \frac{\sum_{i=1}^k n_i x_i}{N} \quad \text{moyenne arithmétique pondérée (k le nombre de classes).}$$

N = nombre d'observations
X_i = valeur de l'observation i
i allant de 1 à N

= somme des valeurs observées
divisée par le nombre d'observations

Tendance centrale

Moyenne (2)

Exemples

1) Les notes (sur 20) à un devoir de français d'une classe de 20 élèves âgés de 10 ans sont les suivantes : 5, 6, 6, 8, 9, 9, 9, 10, 11, 12, 12, 12, 12, 13, 13, 14, 15, 17, 18, 19.

La moyenne arithmétique est égale à



2) Ex : Nombre d'enfants par famille dans un échantillon de 10 familles

2, 8, 1, 1, 4, 5, 2, 3, 0, 1

nombre moyen d'enfants par famille ?



Tendance centrale

Moyenne (2)

Exemples

1) Les notes (sur 20) à un devoir de français d'une classe de 20 élèves âgés de 10 ans sont les suivantes : 5, 6, 6, 8, 9, 9, 9, 10, 11, 12, 12, 12, 12, 13, 13, 14, 15, 17, 18, 19.

La moyenne arithmétique est égale à

$$\frac{5 + 6 + 6 + 8 + 9 + 9 + 9 + 10 + 11 + \dots + 17 + 18 + 19}{20} = 11,5$$

2) Ex : Nombre d'enfants par famille dans un échantillon de 10 familles

2, 8, 1, 1, 4, 5, 2, 3, 0, 1

nombre moyen d'enfants par famille ? **$27 / 10 = 2,7$**

Tendance centrale

Moyenne (3)

Exemples

3) Nombre moyen de voitures des 200 familles enquêtées : Moyenne pondérée

Répartition des 200 familles
selon leur nombre de voitures

Nombre de voitures	Effectifs
0	20
1	100
2	67
3	10
Au moins 4	3

La dernière classe comptabilise le nombre de familles ayant plus de quatre voitures mais ne précise pas le nombre exact de voitures, il est donc nécessaire de faire un choix pour estimer le nombre moyen. Nous pouvons prendre par exemple la décision de considérer que les 3 familles ont 4 voitures, sous cette condition le nombre moyen de voitures par famille est égal à :



Tendance centrale

Moyenne (3)

Exemples

3) Nombre moyen de voitures des 200 familles enquêtées : Moyenne pondérée

Répartition des 200 familles
selon leur nombre de voitures

Nombre de voitures	Effectifs
0	20
1	100
2	67
3	10
Au moins 4	3

La dernière classe comptabilise le nombre de familles ayant plus de quatre voitures mais ne précise pas le nombre exact de voitures, il est donc nécessaire de faire un choix pour estimer le nombre moyen. Nous pouvons prendre par exemple la décision de considérer que les 3 familles ont 4 voitures, sous cette condition le nombre moyen de voitures par famille est égal à :

$$(20*0+100*1+67*2+10*3+3*4)/200 = 1,38$$

Tendance centrale

Quel est le paramètre le plus représentatif des données ?

Ex : Nombre d'enfants dans 10 familles

0, 1, 1, 1, 2, 2, 3, 4, 5, 8

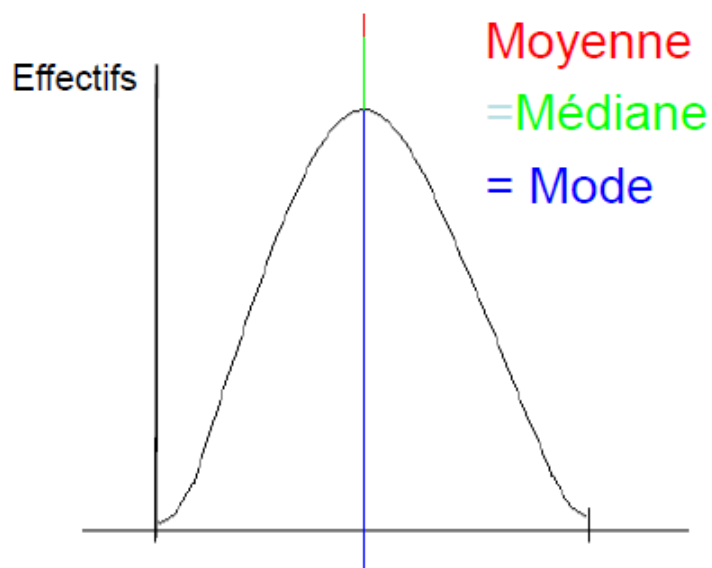
moyenne = 2,7

médiane = 2

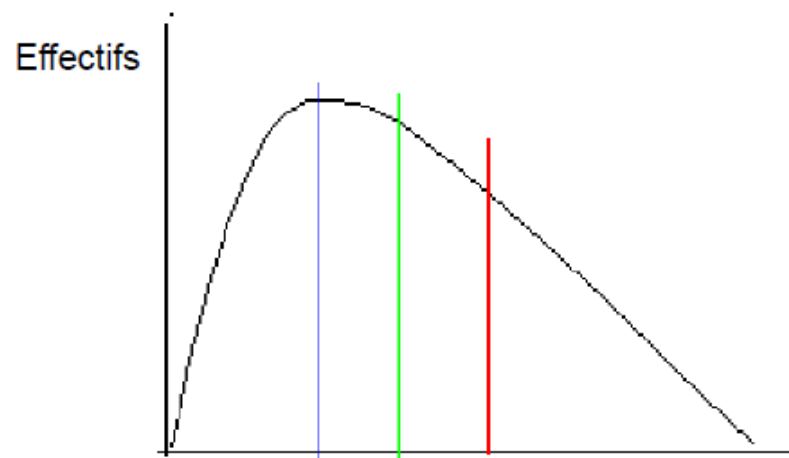
mode = 1

?

Distribution symétrique
(gaussienne, normale)



Distribution asymétrique



Avantages/inconvénients des mesures de tendance centrale des variables quantitatives

Avantages	Inconvénients
<p>Moyenne</p> <ul style="list-style-type: none">▪ Facile à calculer▪ Paramètre connu et répandu, facilement interprétable	<ul style="list-style-type: none">▪ Très influencée par les valeurs extrêmes▪ Représente mal les valeurs d'une population hétérogène (par exemple bimodale) ou fortement asymétrique
<p>Médiane</p> <ul style="list-style-type: none">▪ Peu influencée par les valeurs extrêmes▪ Bon indicateur pour des variables asymétriques	<ul style="list-style-type: none">▪ Plus compliquée à définir (ou calculer) manuellement▪ Ne tient pas compte de l'ensemble des données
<p>Mode</p> <ul style="list-style-type: none">▪ Non influencé par les valeurs extrêmes▪ Permet de présenter des populations hétérogènes qui présentent plusieurs valeurs dominantes	<ul style="list-style-type: none">▪ Compliqué à définir manuellement si N est grand▪ Varie avec la largeur des classes

Tendance centrale

Fréquence d'une catégorie

Valable pour les variables qualitatives

$$\text{fréquence} = \frac{n}{N}$$

← Effectif de la catégorie
← Effectif de la population d'étude

Exemples:

1) La promotion des 6^{ème} A de Médecine de Ouahigouya compte 4 femmes et 25 hommes.

La fréquence des femmes est: $4/(4+25) = 0,1379$ ou 13,8%

2) Parmi les 255 personnes dépistées pour la Covid-19 aux urgences du CHUR de Ouahigouya, 52 étaient des fonctionnaires, 37 des étudiants, 89 des agriculteurs et 77 des commerçants.

La répartition des dépistés pour la Covid-19 selon leur profession est:

	n	%
Profession		
Agriculteur	89	34,9%
Commerçant	77	30,2%
Fonctionnaire	52	20,4%
Etudiant	37	14,5%

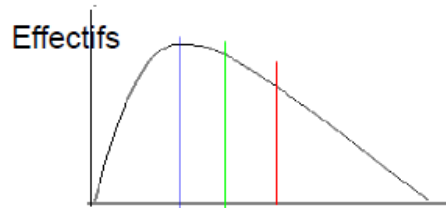
Fréquences

Attention! La somme des fréquences de toutes les catégories = 1 ou 100%

Dispersion

Deux séries peuvent avoir la même moyenne arithmétique et pourtant avoir des dispersions très différentes. Il est donc nécessaire de compléter les informations des caractéristiques centrales par les caractéristiques de dispersion.

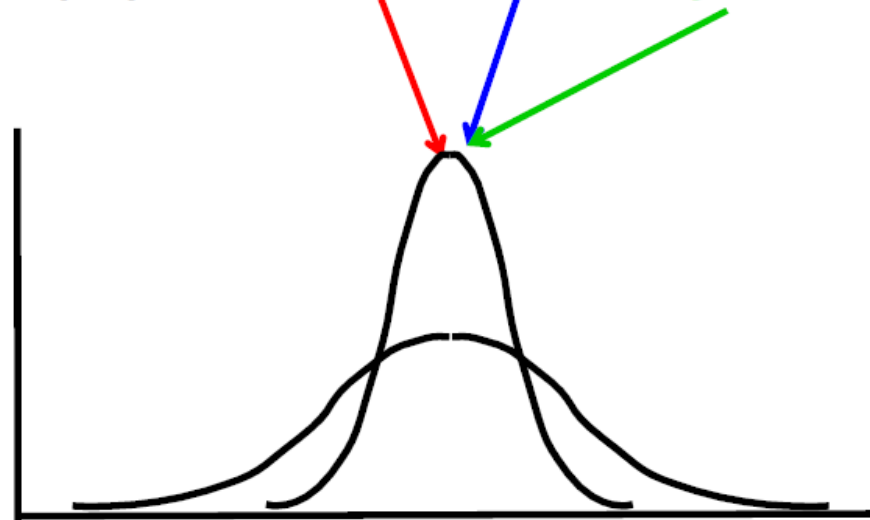
Distribution asymétrique



Tendance centrale identique
MAIS
distribution différente

Distribution symétrique (gaussienne, normale)

superposition **mode** **médiane** **moyenne**

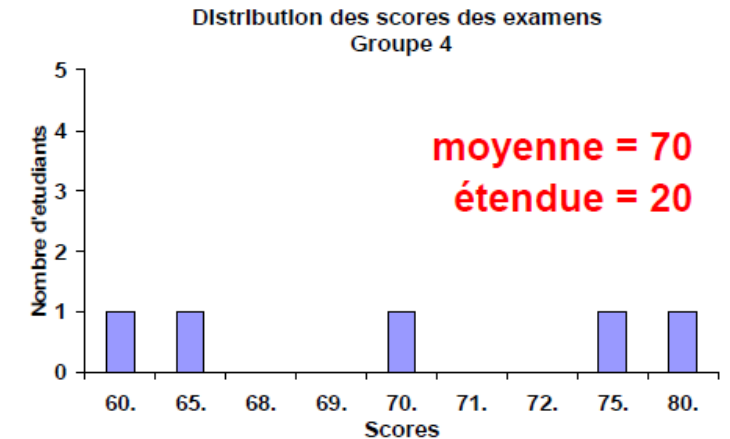
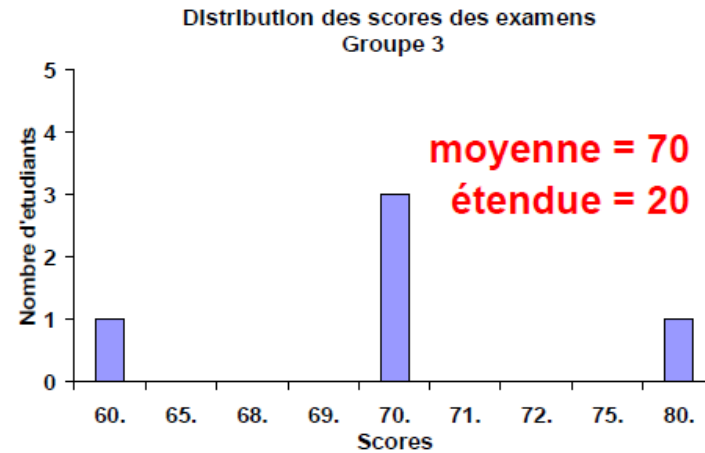
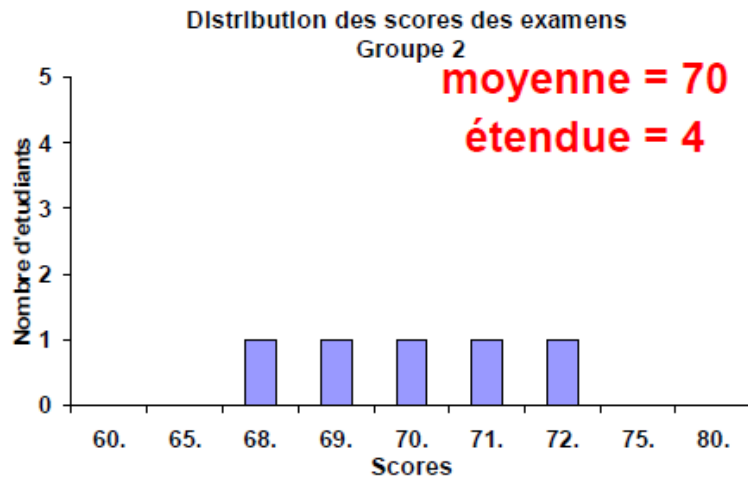


Dispersion

Étendue

Différence entre la valeur la plus élevée et la valeur la plus petite. Au lieu de donner la mesure de l'étendue il est en réalité plus pertinent de donner la valeur minimale et la valeur maximale.

Un inconvénient de cette mesure est qu'elle ne considère que les extrémités.



Dispersion

Variance/ Ecart-type (1)

- La variance (σ^2) exprime la dispersion des observations autour de la moyenne. Elle tient compte de toutes les observations. **Par définition, l'écart moyen par rapport à la moyenne est nul**
- Elle est la somme des distances élevées au carré entre les valeurs et leur moyenne. Cette mesure n'ayant pas la même unité que les données, on utilise la racine carrée de celle-ci qui est appelée écart-type.
 - Variance = moyenne des carrés des écarts à la cette moyenne**
 - Ecart-type = racine carrée de la variance**

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

$$s^2 = \frac{\sum (x_i - m)^2}{n - 1}$$

← Pour un échantillon

σ^2 = variance

x_i = valeur de l'observation i
 i allant de 1 à N

μ = moyenne

N = nombre total d'observations

Dispersion

Variance/ Ecart-type (2)

- Une autre écriture de la formule

variance simple (données brutes)

$$\sigma^2 = \frac{1}{N} \cdot \sum_{i=1}^N (x_i - \bar{X})^2$$

variance pondérée (données regroupées en k

classes)
$$\sigma^2 = \frac{1}{N} \cdot \sum_{i=1}^k n_i \cdot (x_i - \bar{X})^2$$

l'écart-type simple
$$\sigma = \sqrt{\frac{1}{N} \cdot \sum_{i=1}^N (x_i - \bar{X})^2}$$

l'écart-type pondéré
$$\sigma = \sqrt{\frac{1}{N} \cdot \sum_{i=1}^k n_i \cdot (x_i - \bar{X})^2}$$

N = nombre d'observations

X_i = valeur de l'observation i

i allant de 1 à N

n_i = effectif correspondant à la valeur de l'observation

\bar{X} = moyenne

Dispersion

Variance/ Ecart-type (2)

- Une autre écriture de la formule

$$\text{Var}(X) = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{X})^2 = \frac{\sum_{i=1}^N x_i^2 - \frac{(\sum_{i=1}^N x_i)^2}{N}}{N} = \frac{\sum_{i=1}^N x_i^2}{N} - \bar{X}^2$$

$$\text{Var}(X) = \frac{1}{N} \sum_{i=1}^k n_i (x_i - \bar{X})^2 = \frac{\sum_{i=1}^k n_i x_i^2 - \frac{(\sum_{i=1}^k n_i x_i)^2}{N}}{N} = \frac{\sum_{i=1}^k n_i x_i^2}{N} - \bar{X}^2$$

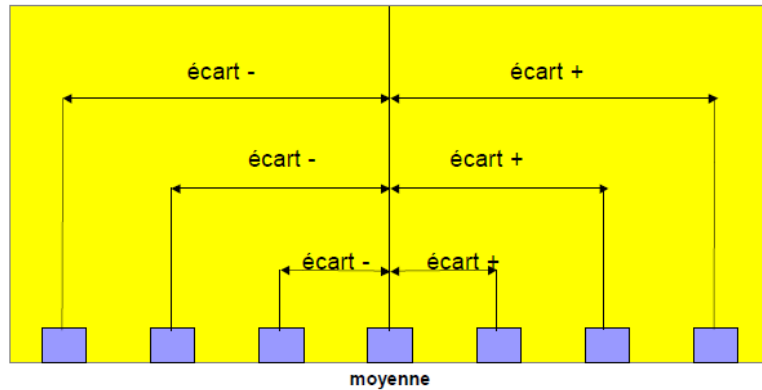
N : Effectif total ; n_i : Effectif de la classe ; k : Nombre de classes

Dispersion

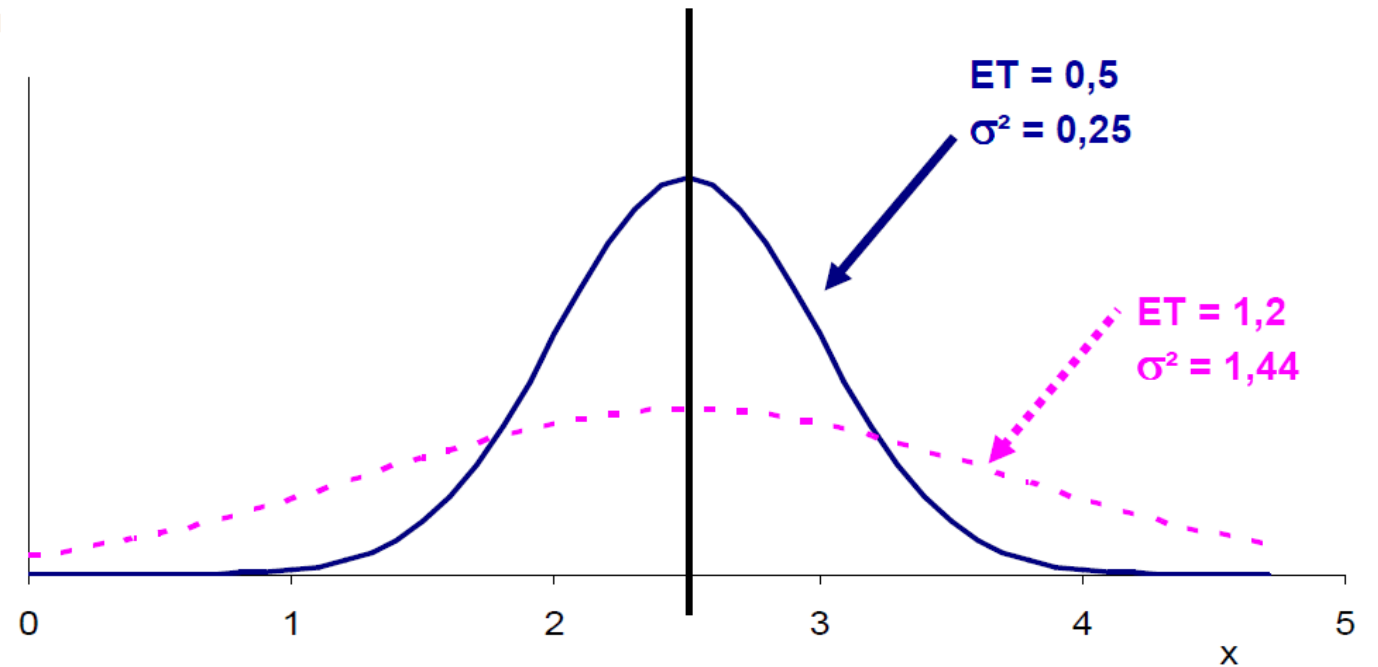
Variance/ Ecart-type (3)

Exemples d'illustrations

- Par définition, l'écart moyen par rapport à la moyenne est nul



Comparaison de 2 distributions de même moyenne (2,5)



Dispersion

Variance/ Ecart-type (4)

Exemples

Les notes (sur 20) à un devoir de français d'une classe de 20 élèves âgés de 10 ans sont les suivants :

5 – 6 – 6 – 8 – 9 – 9 – 9 – 10 – 11 – 12 – 12 – 12 – 12 – 13 – 13 – 14 – 15 – 17 -18 – 19

La valeur de la variance de la valeur de la note est égale à :

Dispersion

Variance/ Ecart-type (4)

Exemples

Les notes (sur 20) à un devoir de français d'une classe de 20 élèves âgés de 10 ans sont les suivants :

5 – 6 – 6 – 8 – 9 – 9 – 9 – 10 – 11 – 12 – 12 – 12 – 12 – 13 – 13 – 14 – 15 – 17 -18 – 19

La valeur de la variance de la valeur de la note est égale à :

$$((5-11,5)^2 + 2*(6-11,5)^2 + \dots + (19-11,5)^2)/20 = 14,45$$

Dispersion

Percentiles

- Valeurs pour lesquelles un certain pourcentage de donnée a une valeur inférieure.
- N^{ième} percentile : valeur en dessous de laquelle se situent n % des observations
- Percentiles les plus utilisés
 - **Quartiles**
 - Définition : valeur qui divise la population d'étude en **4 parties égales**
 - Q1 = 25%, Q2 = 50%, Q3 = 75%
 - **Terciles**
 - Définition : valeur qui divise la population d'étude en **3 parties égales**
 - T1 = 33%, T2 = 66%
 - **Déciles**
 - Définition : valeur qui divise la population d'étude en **10 parties égales**
 - D1 = 10%, D2 = 20%,

NB: La description en percentiles offre une bonne visibilité de la répartition des données

Dispersion

Quartiles

- Premier quartile (Q1) ou 25^{ième} percentile
 - 25% de la population a une valeur de la variable inférieure à celle-ci
 - Calcul: si le 25ème percentile appartient à un intervalle $[a, b[$:

$$a + \frac{(b - a)}{n} * \left(\frac{N}{4} - \sum_{x_{i+1} < a} n_{[x_i; x_{i+1}]} \right)$$

- Deuxième quartile (Q2) = médiane; (déjà vu)

- Troisième quartile (Q3) ou 75^{ième} percentile
 - 75% de la population a une valeur de la variable inférieure à cel
 - Calcul: si le 25ème percentile appartient à un intervalle $[a, b[$:

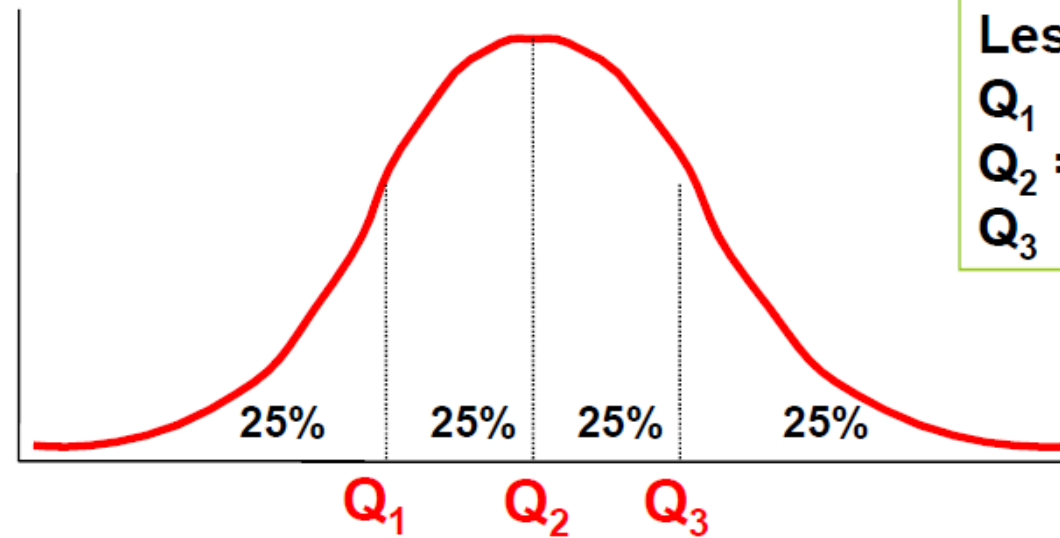
$$a + \frac{(b - a)}{n} * \left(\frac{3 * N}{4} - \sum_{x_{i+1} < a} n_{[x_i; x_{i+1}]} \right)$$

- Intervalle inter-quartile
 - Compris entre Q1 et Q3
 - Correspond à l'intervalle à l'intérieur duquel se situent 50% des données centrales.

Dispersion

Percentiles

La distribution peut être découpée en 4 groupes
→ quartiles



Les quartiles :

Q_1

$Q_2 = \text{médiane}$

Q_3

$Q_3 - Q_1 = \text{intervalle inter quartiles}$

Dispersion

Intervalle de confiance à 95% d'une fréquence

- Renseigne sur la dispersion des observations autour d'une fréquence
- Son calcul obéit à la même logique que celui de celui d'une moyenne

Ecart type d'un pourcentage $\sigma_p \approx \sqrt{\frac{f(1-f)}{n}}$

Intervalle de confiance d'un pourcentage à 95%

$$f - 2 \sqrt{\frac{f(1-f)}{n}} < P < f + 2 \sqrt{\frac{f(1-f)}{n}}$$

n : taille de l'échantillon

f : fréquence dans l'échantillon

2 : approximation de **1,96** = $U_{(0,05)}$

II.3. Tableaux

Tableaux

Pourquoi?

- La rédaction des effectifs ou des fréquences obtenus pour chaque modalité, présentée en texte est souvent longue et peu lisible
- Le tableau permet une **présentation complète et précise des données**
- Il doit toujours être accompagné d'un **titre informatif**, c'est à dire donnant suffisamment d'informations sur le groupe étudié, le lieu et la période de l'étude
- Par convention ce titre doit être **placé au-dessus du tableau**

Répartition selon le sexe des 402 sujets âgés de plus de 18 ans ayant participé à l'enquête « hypertension artérielle » du 19 au 21 avril 1999 à l'Ile Maurice

Sexe	Effectif	Fréquence en pourcentage
Femmes	241	60
Hommes	161	40
Total	402	100

Tableaux

Règles générales de présentation

- Un bandeau de titre pour indiquer la nature des informations figurant dans les colonnes,
- Ce bandeau a un trait horizontal au-dessus et au-dessous. La tête de colonne permet d'indiquer la nature de la variable figurant dans cette colonne,
- Un trait horizontal figure sous la dernière ligne
- aucun autre trait n'est utile, en particulier aucun trait vertical
- Les chiffres sont alignés par colonne : sur la droite (s'il s'agit d'entiers) ou sur la virgule (s'ils sont exprimés avec une décimale)
- Pour une même variable, le même nombre de décimales est employé
 - Par exemple, on ne présenterait pas une proportion de 14,4% pour une catégorie et 9% pour une autre. On choisirait 14% et 9% ou 14,4% et 9,0%
- Les totaux, s'il y a lieu, doivent être donnés
- Les unités de mesure doivent systématiquement être indiquées pour les variables quantitatives. Elles doivent figurer une seule fois à côté du nom de la variable
- En français, le séparateur décimal est la virgule, et non le point utilisé dans le système anglo-saxon.

Répartition selon l'âge des 402 sujets âgés de plus de 18 ans ayant participé à l'enquête « hypertension artérielle » du 19 au 21 avril 1999 à l'Ile Maurice

Tranche d'âge (ans)	Effectif	Fréquence en pourcentage	Fréquences cumulées en pourcentage
[18-28[66	16,4	16,4
[28-38[79	19,7	36,1
[38-48[89	22,1	58,2
[48-58[58	14,4	72,6
[58-68[54	13,4	86,1
[68-78[41	10,2	96,3
[78-88[14	3,5	99,8
≥ 88	1	0,2	100,0
Total	402	100,0	

II.3. Graphiques ou figures

Graphiques

Pourquoi ?

- Vient en complément des informations données dans le texte et les tableaux
- Met à la disposition du lecteur un document qui synthétise l'information
- Aide à visualiser l'information pertinente à retenir
 - Met en évidence **un phénomène remarquable** : contrastes ou tendances
- Le choix de réalisation d'un graphique repose essentiellement
 - le type de variable étudié
 - Le type d'information ou de contraste que l'on souhaite montrer

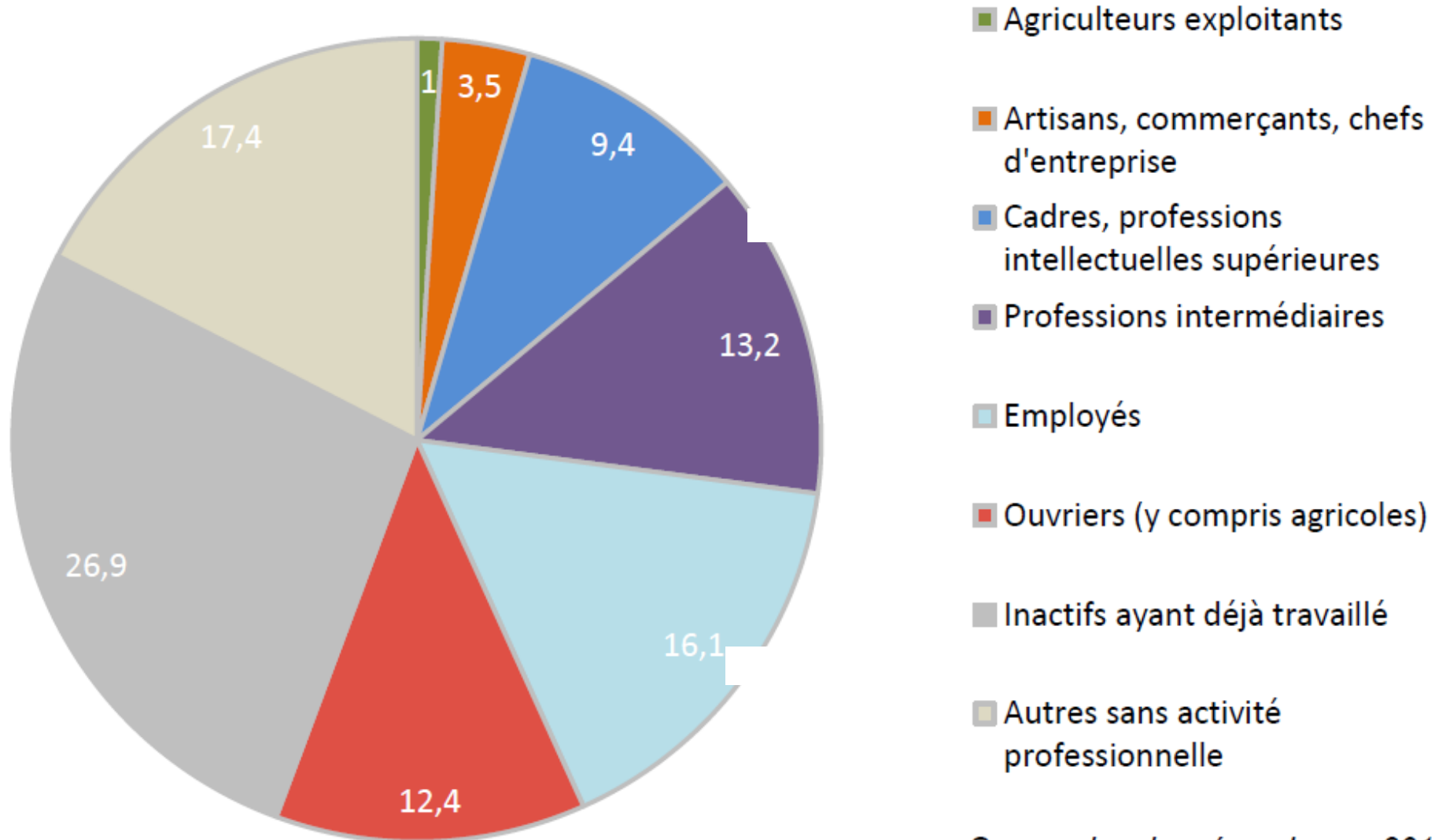
Variables qualitatives

Diagramme en secteur ou Camembert

- Utilisé pour représenter la répartition de variables nominales
- Consiste à partager la surface d'un disque proportionnellement à la surface de chaque modalité

Camembert ou diagramme en secteur

Répartition par CSP de la population française en 2011

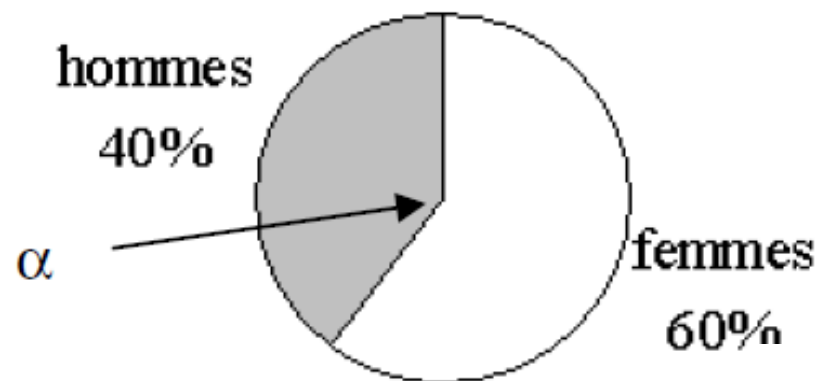


Source des données : Insee, 2011.

Camembert ou diagramme en secteur

Fil rouge : Représentation de la répartition selon le sexe

La répartition selon le sexe des 402 sujets âgés de plus de 18 ans ayant participé à l'enquête « hypertension artérielle » du 19 au 21 avril 1999 à l'île Maurice sera construite afin que les surfaces représentant la proportion de femmes et la proportion d'hommes dans l'étude soient respectivement 60% et 40% de la surface du disque. (l'angle α est proportionnel à la fréquence d'hommes et représente donc ici 40% de 360° , donc $\alpha = 0,40 \times 360 = 144^\circ$)



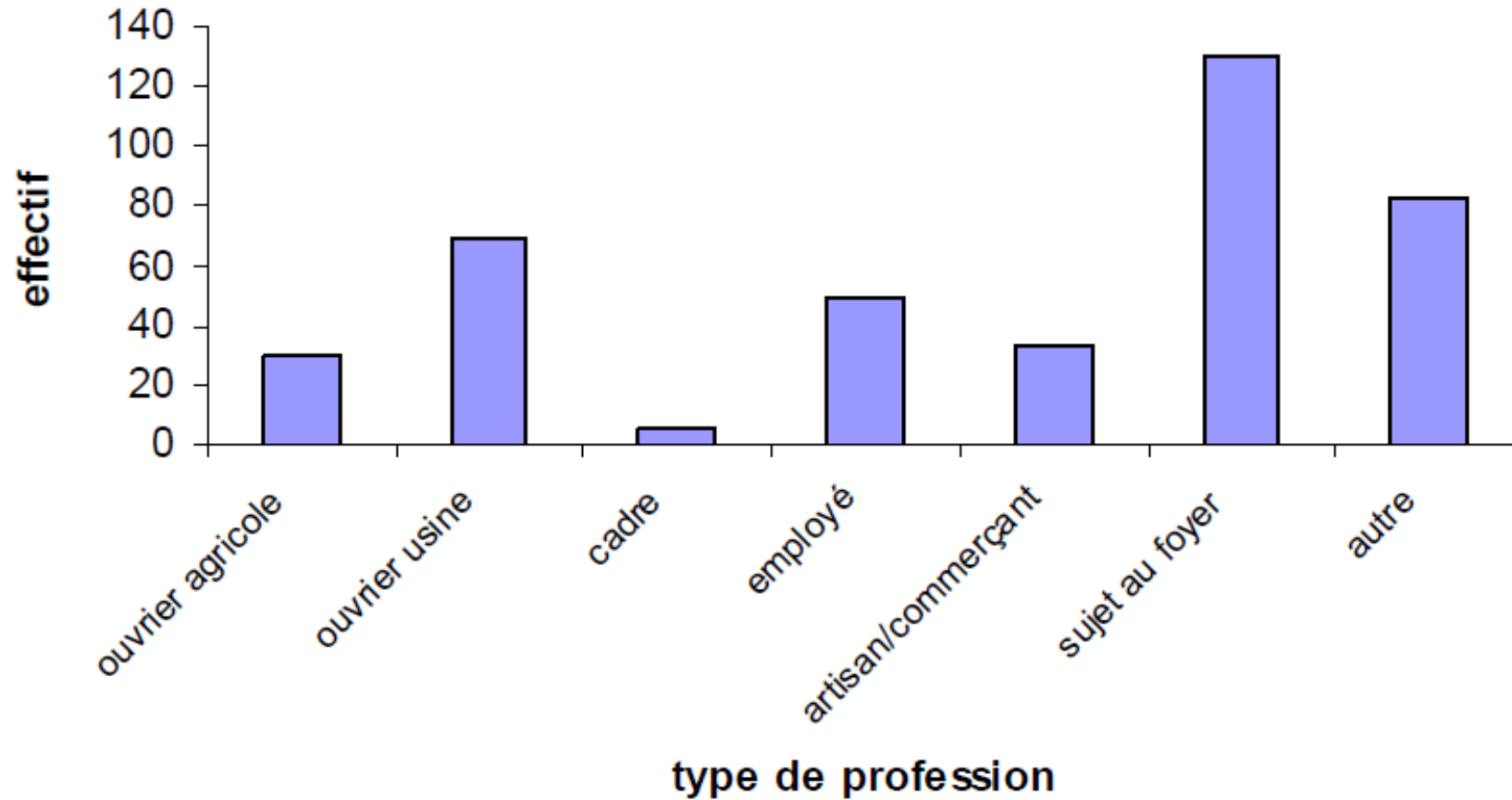
Graphique 1 : Répartition selon le sexe des 402 sujets âgés de plus de 18 ans ayant participé à l'enquête « hypertension artérielle » du 19 au 21 avril 1999 à l'île Maurice

Variables qualitatives

Diagramme en barres ou en bâtons

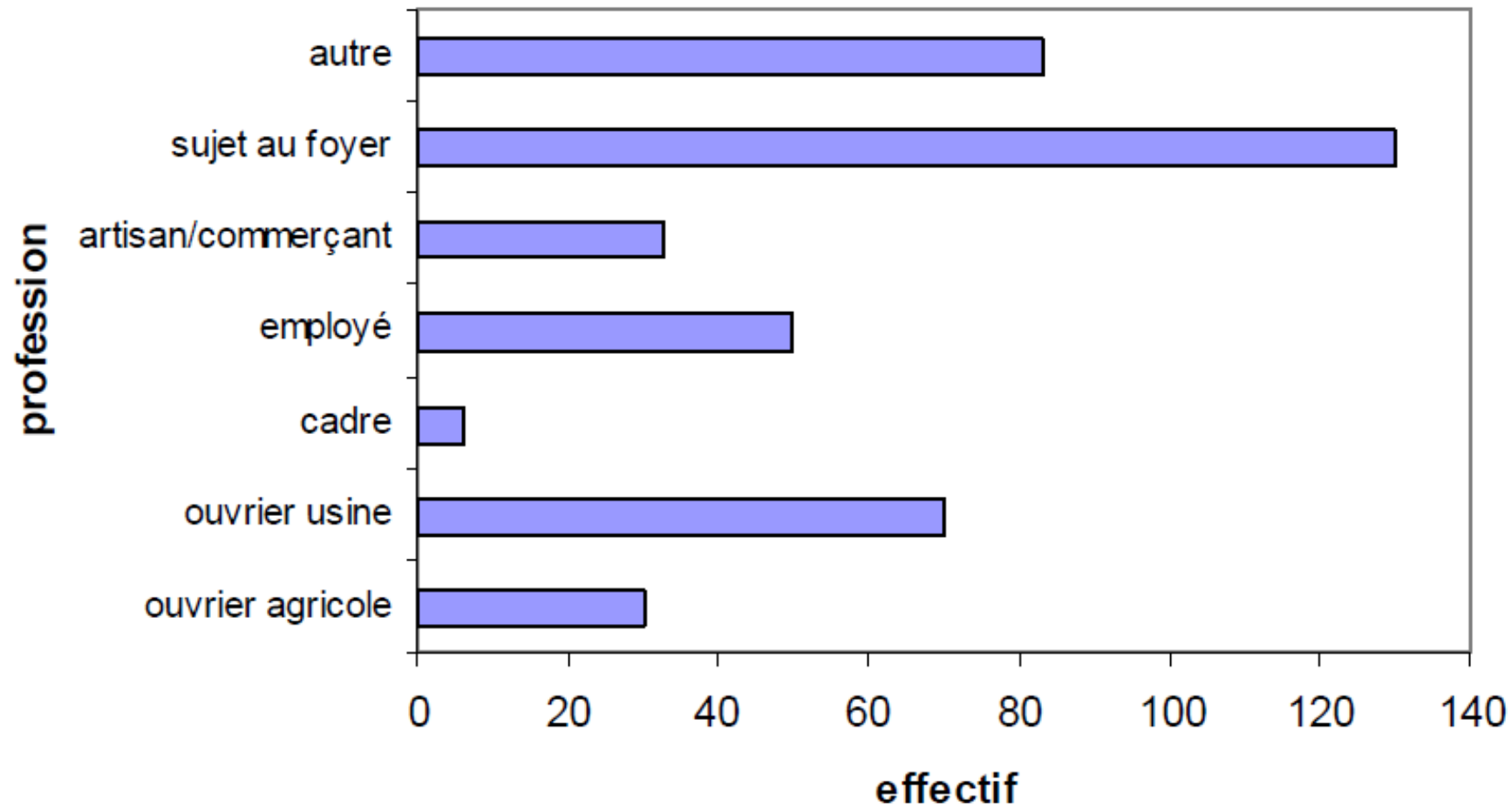
- Consiste à représenter l'effectif ou la fréquence de chaque modalité en répertoriant sur l'axe horizontal, appelé **axe des abscisses**, le nom des modalités et sur l'axe vertical, appelé **axe des ordonnées** les effectifs, ou les fréquences, de chaque modalité.
- La représentation peut aussi être inversée pour une meilleure lisibilité.

Diagramme en barres ou en bâtons



Graphique 3 : Répartition selon la profession des 102 sujets âgés de plus de 18 ans ayant participé à l'enquête « hypertension artérielle » du 19 au 21 avril 1999 à l'île Maurice

Diagramme en barres ou en bâtons



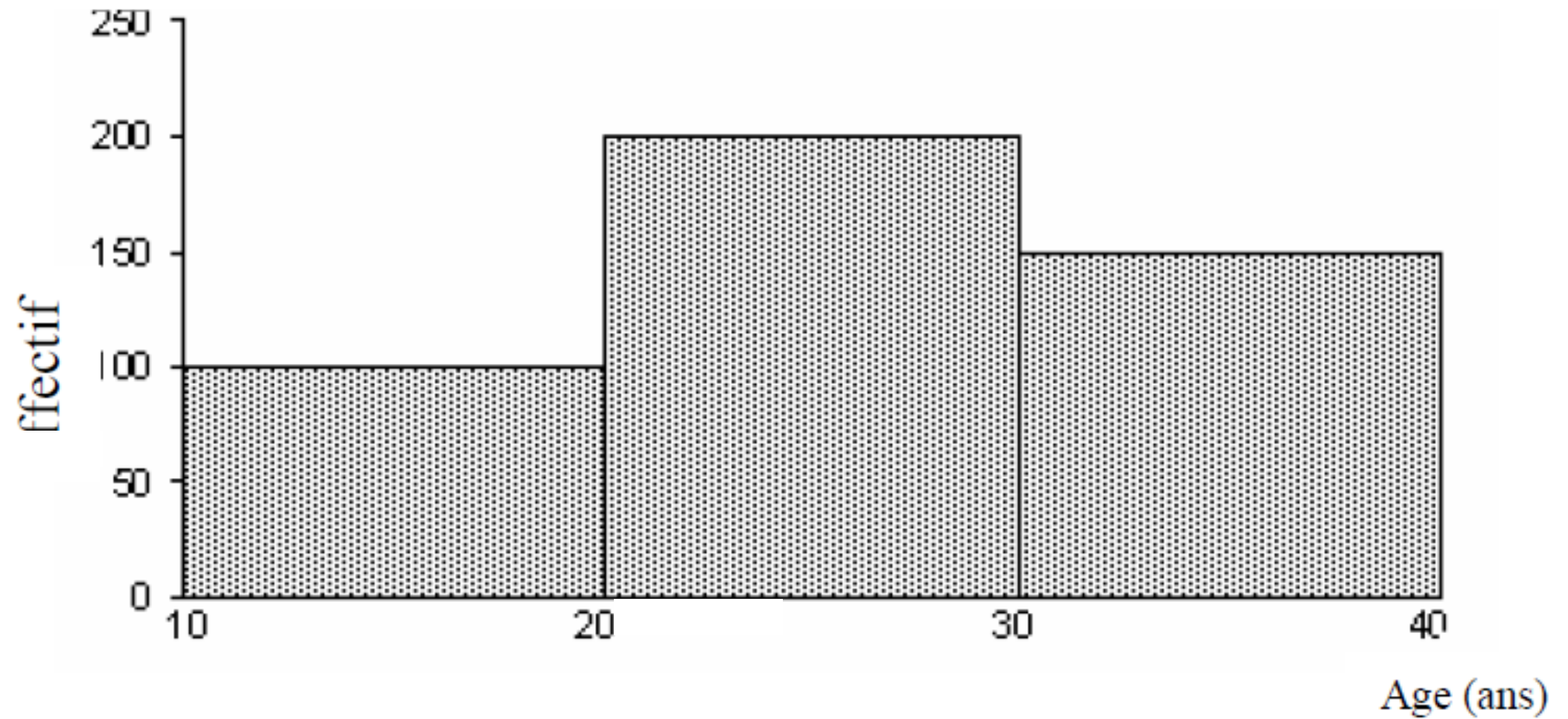
Graphique 4 : Répartition selon la profession des 92 sujets âgés de plus de 18 ans ayant participé à l'enquête « hypertension artérielle » du 19 au 21 avril 1999 à l'île Maurice

Variables quantitatives continues

Histogramme

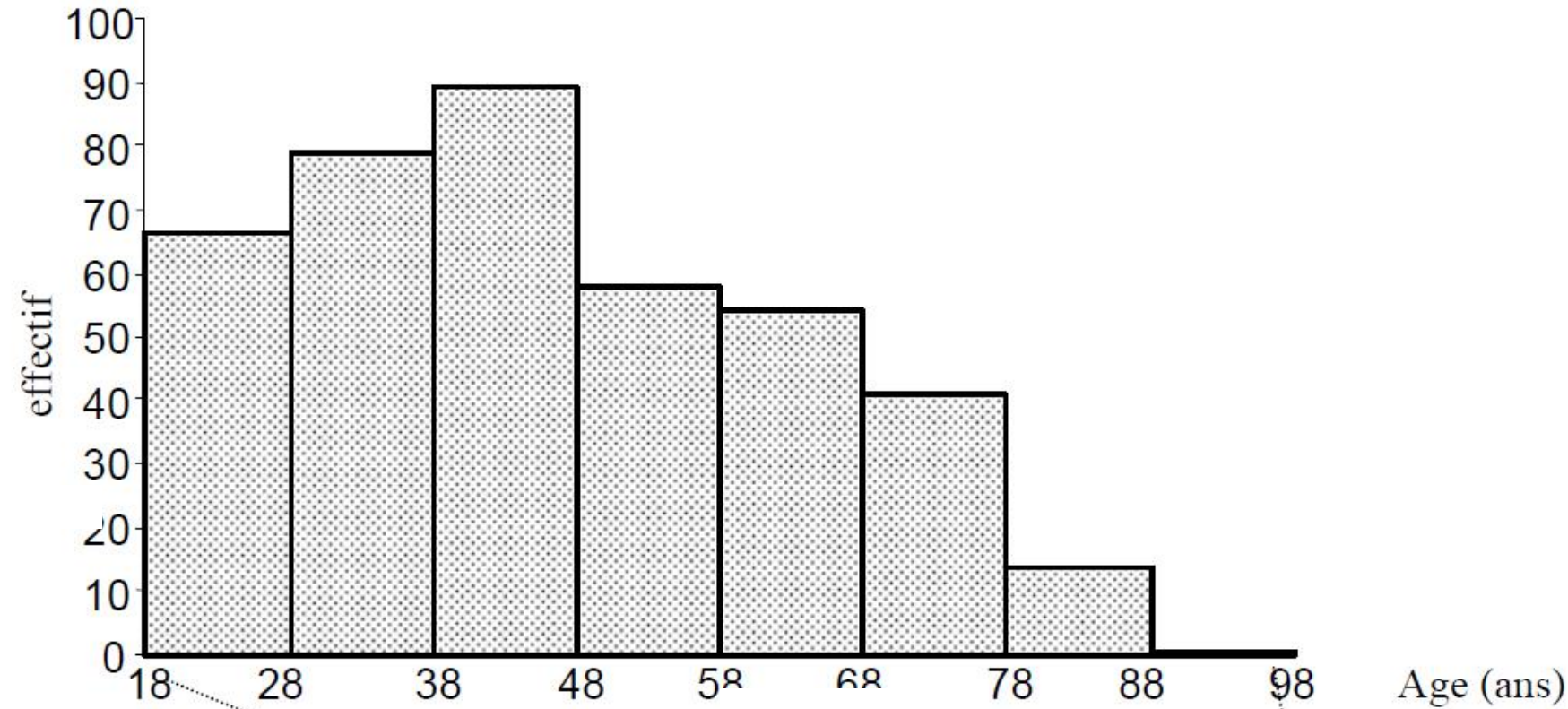
- Les valeurs des bornes minimales et maximales de la distribution doivent être connues et fixées à l'avance
- Consiste à représenter des rectangles accolés les uns aux autres
- La surface de chaque rectangle devant être proportionnelle à l'effectif des sujets compris dans l'intervalle
- il est nécessaire de définir l'amplitude des intervalles de regroupement, puis de déterminer la hauteur du rectangle pour conserver cette proportionnalité
- Afin de simplifier la lecture du graphique il est fortement conseillé d'utiliser des intervalles de même amplitude

Exemple : une étude réalisée auprès de 400 sujets ayant un âge compris entre 10 et 40 ans : 100 personnes ont une valeur comprise dans l'intervalle $[10 ; 20[$, 200 dans l'intervalle $[20 ; 30[$ et 150 dans l'intervalle $[30 ; 40[$. Le second rectangle sera 2 fois plus haut que le premier et le troisième une fois et demi, car la base du rectangle est de même longueur.



Répartition de 400 sujets selon leur âge.

La répartition de l'âge est représentée en utilisant des intervalles d'amplitude de 10 ans.



Répartition des 402 sujets âgés de plus de 18 ans ayant participé à l'enquête « hypertension artérielle » du 19 au 21 avril 1999 à l'île Maurice selon leur âge

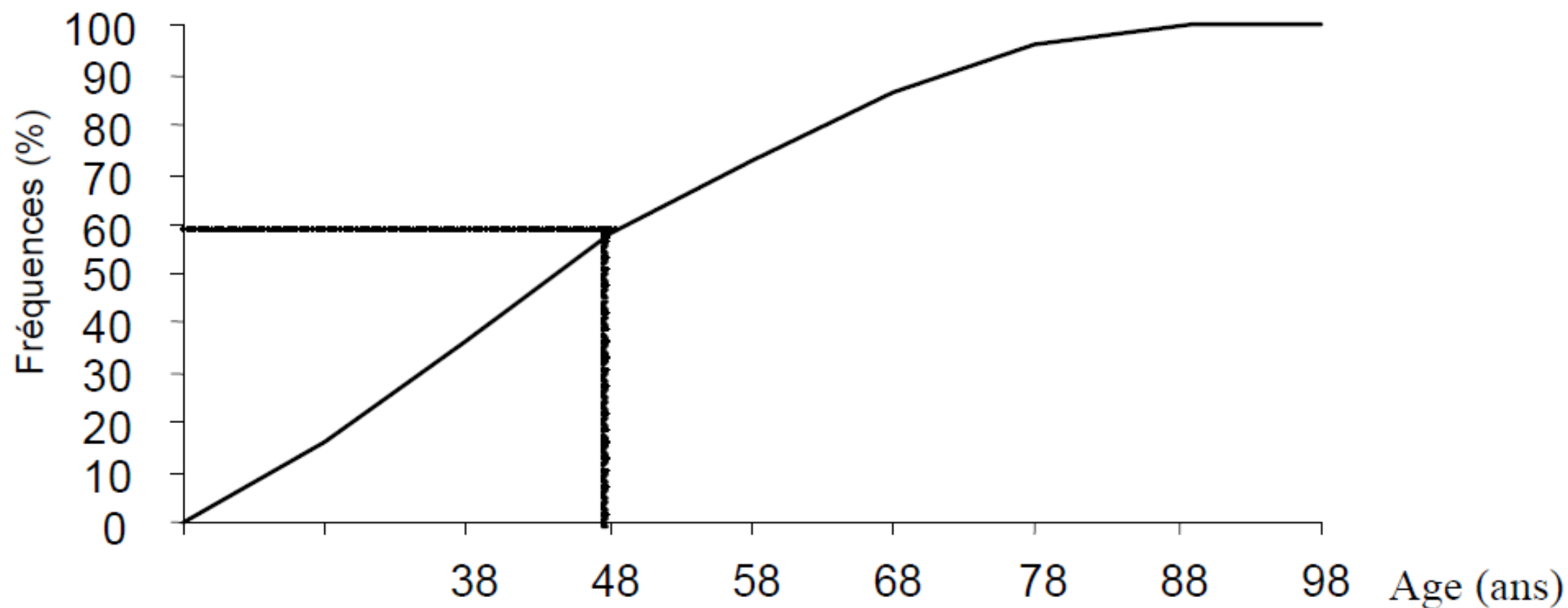
Commentaire : Les 402 sujets sont âgés de plus de 18 ans et de moins de 98 ans. 89 d'entre eux ont au moins 38 ans et moins de 48 ans, c'est à dire entre 38 et 48 ans.

Variables quantitatives continues

Polygones des fréquences cumulées

- la représentation graphique qui joint par des segments de droite les points ayant pour abscisse la borne supérieure de l'intervalle et pour ordonnée la fréquence cumulée correspondant à cet intervalle.
- Ce polygone a donc pour point de départ l'origine du graphique et pour point d'arrivée le point ayant pour abscisse la valeur maximale de la variable étudiée et pour ordonnée la valeur 1 (ou 100%) si l'on étudie les fréquences cumulées, ou la valeur n (effectif total) si on travaille avec les effectifs cumulés
- le polygone cumulé a la même forme qu'il représente les fréquences cumulées ou les effectifs cumulés.

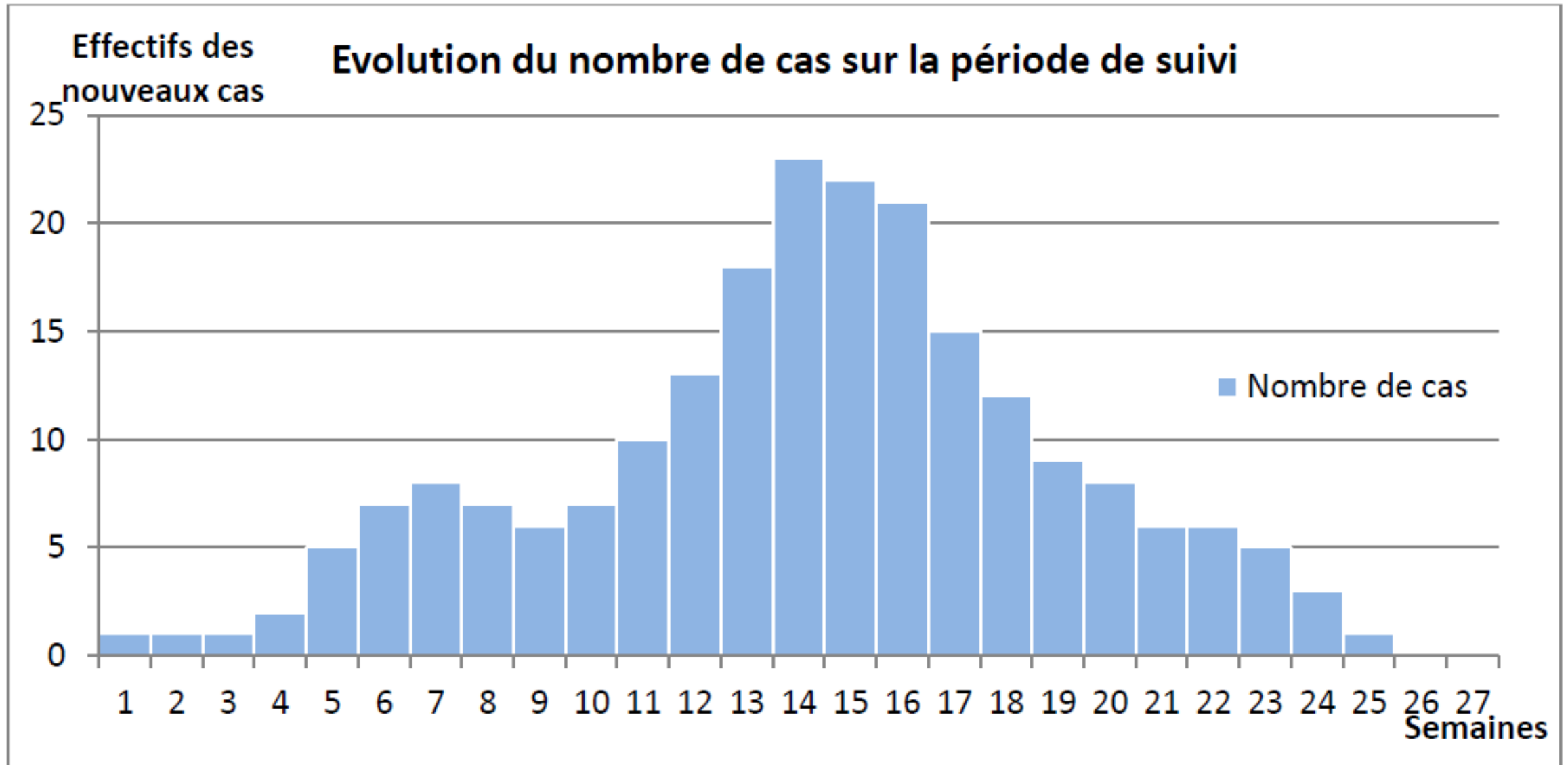
NB: Ce type de figure peut également utilisé pour décrire une variable ordinale



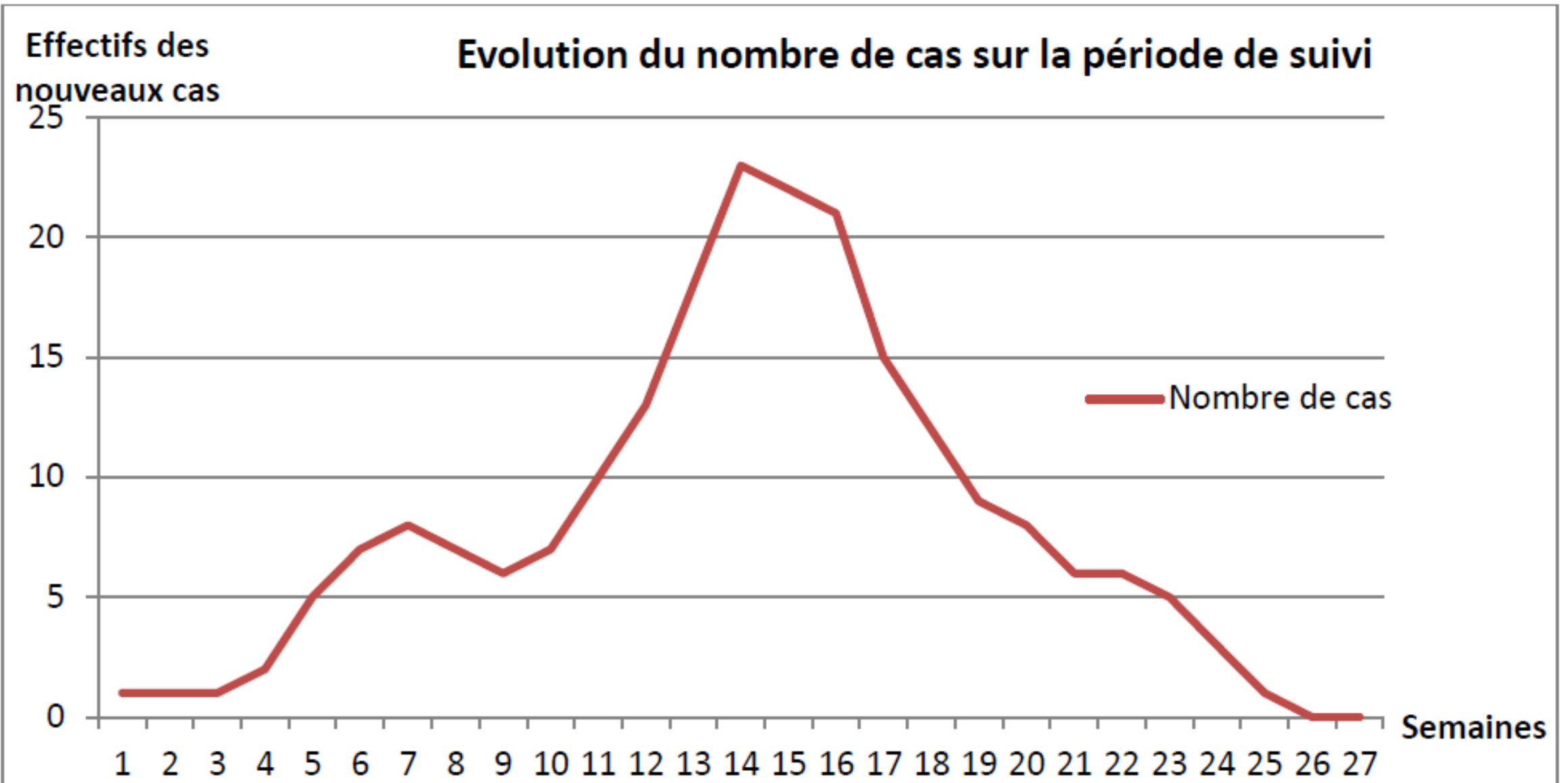
Répartition des 402 sujets âgés de plus de 18 ans ayant participé à l'enquête « hypertension artérielle » du 19 au 21 avril 1999 à l'Ile Maurice selon leur âge. Représentation par un polygone de fréquences cumulées

Commentaire : Un peu moins de 60% des 402 sujets ont moins de 48 ans.

Variables quantitatives discrètes



Variables quantitatives discrètes



Merci pour votre attention

Documents ressources

1. François Dabis, Jean Claude Desenclos. Epidémiologie de terrain. Méthodes et applications 2017
2. Institut National de Veille Sanitaire. Recommandations pour la représentation des résultats au DES. Juillet 2015
3. Jean Bouyer. Méthodes statistiques. Médecine-Biologie. 2017.
4. Thierry Ancelle. Statistique-Epidémiologie, 4^{ème} édition. 2017.