Chap1 : NOTIONS DE BASE DE L'ANALYSE DE DONNEES

I- DEFINITIONS ET EXEMPLES:

Les données sont des informations issues d'observations, de mesures faites sur une population humaine, animale ou de chose (équipement et matériel physique ou logique).

En analyse de donnée, on s'accorde sur quelques définitions :

Population et individus : La population est l'ensemble des individus qui nous intéresse dans l'étude (les individus sont aussi appelés unités statistiques). La taille, d'une population est en général désignée par **N**.

Variable (ou caractère) valeurs : Une variable est une information dont on recueille par l'observation ou la mesure sur *chaque* individu.

On parle de variable parce que la valeur de l'information n'est pas la même d'un individu à l'autre.

Effectif : Nombre d'individus, d'une population ou d'une partie quelconque de cette population.

Fréquence (proportion de la population) : Rapport d'un effectif donné d'individus à la taille de la population.

Recensement : Recueil des valeurs de la totalité des individus d'une population donnée. **Sondage, *n*-échantillon, base de sondage, taux de sondage :** Un sondage consiste à procéder au recueil des valeurs d'une partie ou encore d'un **échantillon**) d'effectif *n* (d'où l'expression *n*-échantillon) de la population d'effectif N. (***base de sondage***). Le taux de sondage est le rapport *n*/**N**.

Variable qualitative (ou nominale) : Variable dont les valeurs (ou modalités) observées sont telles qu'il est impossible d'attribuer une valeur unique à la réunion de deux (ou plusieurs) individus par une opération mathématique sur leurs valeurs. Exemple du "statut matrimonial". Les valeurs observées peuvent néanmoins être numériques.

Variable, ou caractère ordinal : Variable qualitative dont on peut tout de même comparer les modalités entre elles, et, par conséquent, ranger par "valeurs croissantes" ou "décroissantes". Exemple de l'appréciation d'un produit par des utilisateurs.

Variable, ou caractère quantitatif : Variable *numérique* pour laquelle on peut, faire une opération mathématique quelconque, comme l'addition, la multiplication, pour deux (ou plusieurs) individus.

Exemple 1 :

Sur chacun des individus sondés, on observe un **caractère** (ou **variable**) ci-après:

- âge
- sexe
- revenus
- métier
- nombre d'enfants
- pression artérielle
- durée de bon fonctionnement,
- fumeur
- titulaire du permis B
- ingénieur
- docteur

Solution : Ce caractère est **quantitatif** s'il est possible de le mesurer, donc de le représenter avec un nombre :

- âge
- revenus
- nombre d'enfants
- pression artérielle
- durée de bon fonctionnement
- ingénieur
- docteur

Il est **qualitatif** dans le cas contraire :

- métier

- sexe
- fumeur
- titulaire du permis B
- ingénieur
- docteur

Une valeur prise par une variable s'appelle une **modalité**.

Exemple 2 :

1- La variable statistique "couleur de téléphone portable" est-elle :
   a- qualitative
   b- quantitative
   c- discrète
   d- continue
2- La variable statistique " salaire brut" est-elle :
   a- qualitative
   b- quantitative
   c- discrète
   d- continue
3-La variable statistique "nombre de machine réparées" est-elle :
   a- qualitative
   b- quantitative
   c- discrète
   d- continue
4- on donne les variables suivants :
Hauteur, Poids, Rendement, Chiffre d'affaire, Cylindrée, Marge de puissance, Affaiblissement en dB de signal, Rapport signal sur bruit.
   a- Montrer le caractère quantitatif de ces variables
   b- Préciser les modalités qui peuvent transformer l'étude quantitative de ces variables en et de qualitatives

Solution : 1) Pour le premier cas, la variable statistique est qualitative. 2) Pour le deuxième cas, la variable statistique est quantitative continue. 3) Pour le troisième cas, la variable statistique est quantitative discrète

Solution question 4 (cf. tableau ci-après) :

Variable quantitative	Modalités qualitatives envisageables	commentaires
hauteur	Petit, Moyen, Grand	
poids	Très léger, Léger, Moyen, Lourd, Très lourd	
rendement	Faible, Moyen, Elevé	
Chiffre d'affaire	Modéré, Moyen, Important, Très important	
cylindrée	Petite, Moyenne, Grosse	
Marge de puissance,		

Affaiblissement en dB de signal,		
Rapport signal sur bruit.		

II- ECHANTILLONNAGE

La façon de sélectionner l'échantillon est aussi importante que la manière de l'analyser.

- l'échantillon **doit être représentatif de la population** (l'échantillonnage aléatoire est le meilleur moyen d'y parvenir)
- Un **petit échantillon représentatif est, de loin, préférable** à un grand échantillon biaisé.
-

ECHANTILLONNAGE ALEATOIRE SIMPLE

Un *échantillon aléatoire* est un échantillon *tiré au hasard* dans lequel tous les individus ont la *même chance* de se retrouver ; dans le cas contraire, l'échantillon est *biaisé*.

L'échantillonnage aléatoire simple est à la base de l'ensemble de la théorie d'échantillonnage.

Pour obtenir un échantillon par cette méthode, il faut numéroter les individus de la population de 1 à N, puis on tire n individus. Le tirage est généralement réalisé sans remise.

L'objectif étant de fournir une estimation sans biais de la moyenne et de la variance de la population.

ECHANTILLONNAGE PAR GRAPPES

Pour l'**échantillonnage par grappes**, **il faut subdiviser la population en sous-ensembles aussi appelées grappes.**

Chacun de ces sous-ensembles **doit être représentatif de la population source (ou population mère).**

L'échantillonnage par grappes consiste donc à tirer aléatoirement des individus au sein des grappes choisies et mener l'étude sur ces individus.

Par Exemple une ville, subdivisée en quartiers ; Un échantillonnage aléatoire simple est alors effectué.

Pour avoir un échantillonnage par grappes ayant les propriétés précises **qu'il faut veiller à ce que la taille des grappes soit uniforme**. Il faut aussi **avoir une homogénéité des individus** composant les grappes.

ECHANTILLONNAGE PAR LA METHODE DES QUOTAS

La **méthode des quotas** est basée sur la **répartition en données connues d'une population telles que :** âge, sexe, situation géographique, catégorie socio-professionnelle.

Une fois la dimension du sondage que l'on souhaite effectuer, il suffit de calculer le nombre d'individus par chaque critère choisi.

Cette méthode repose sur l'hypothèse que l'information que l'on souhaite obtenir est corrélée avec la population.

Exemple 3 *:*

Nous désirons déterminer la capacité moyenne des réseaux métropolitains. On a des réseaux internet métropolitains (RIM) et de réseaux téléphoniques Métropolitains (RTM) en Mbps à partir d'un échantillon de 10 réseaux (Pour le nombre total d 86 réseaux la capacité moyenne est de 174,0 Mbps).
Mus par une bonne intention, sachant que les (RTM) sont, en général, plus grands que les (RIM), nous choisissons un échantillon contenant autant de (RIM) que de (RTM) .
Soient 5 RIM et 5 RTM choisis au hasard :

RIM (Mbps)	RIM (Mbps)
171	193
165	187
173	180
174	185
166	178

Solution :
A partir de cet échantillon de 10 individus, nous obtenons une taille moyenne des reseaux de 177,2 Mbps, soit 3,2 Mbps de plus que la valeur exacte.

Avons-nous procédé correctement au choix de l'échantillon, sachant que la population contient 51 RIM et 35 RTM (86 réseaux étudiés) ?

Non, car chaque RTM avait plus de chances d'être choisi que chaque RIM.

En effet, les 5 RTM étant tirés au hasard dans une population de 35 individus, chacun d'eux avait 5 chances sur 35 d'être choisi, soit une probabilité de 5/35 ≅ 0,143.

Les 5 RIM étant choisies dans une population de 51 individus, chacun a 5 chances sur 51 d'être choisi, soit une probabilité de 5/51 ≅ 0,098, donc nettement plus faible que pour les RTM.

Nous avons biaisé l'échantillon en faveur des RTM. Il n'est donc pas surprenant que nous obtenions un résultat trop élevé.

La manière correcte de procéder est de choisir au hasard dans toute la population, sans considération du sexe.

Un tel tirage au hasard a donné les tailles suivantes de réseaux (en Mbps) :

187, 165, 180, 168, 165, 160, 174, 183, 168, 176

La moyenne de l'échantillon est de 172,6 Mbps.

Elle est plus proche de la valeur exacte (erreur de – 1,4 mbps).

[En fait, vu les petits échantillons utilisés, le hasard aurait pu donner un résultat inverse. Ce sera beaucoup moins probable pour de grands échantillons. Le raisonnement est néanmoins valable en toute généralité].

Une autre manière de procéder est d'utiliser la technique des *quotas*.

Sachant que la population étudiée contient 35/86 ≅ 40% de RTM et 51/86 ≅ 60% de RIM, nous pourrions nous assurer que l'échantillon respecte les mêmes proportions, soient 4 RTM et 6 RIM.

Exemple 4:

Les échantillons suivants sont-ils représentatifs de la population visée ?

1. Pour connaître les opinions politiques de la population d'une ville, on envoie 5 enquêteurs pour interroger les gens à la sortie de 5 grands magasins. Ils doivent questionner les clients jusqu'à ce qu'ils réunissent, chacun, un échantillon de 200 réponses.

2. On désire faire une enquête sur les réseaux sociaux les plus utilisés. Pour cela, on choisit au hasard 1000 numéros de téléphone dans l'ensemble des annuaires et on les appelle pendant les heures de bureau. On obtient 583 réponses.

Solution :

1- Non, car les clients des supermarchés ne sont pas typiques de l'ensemble de la population (en général, dans un ménage, c'est toujours la même personne qui fait les courses; l'échantillon contiendra probablement trop de femmes, d'inactifs,…).

2- Non car cet échantillon élimine pratiquement tous les individus actifs (étudiants, travailleurs, …).

Une amélioration de cet échantillon consisterait à téléphoner en soirée et à répéter l'appel pendant plusieurs jours si on n'obtient pas de réponse, de telle manière que l'échantillon obtenu se rapproche le plus possible de l'échantillon sélectionné.

III- PARAMETRES STATISTIQUES D'UNE SÉRIE DE DONNEES

Une série de données peut se caractériser par 2 grands types de paramètres:

Paramètres de position : ils donnent l'ordre de grandeur des observations et sont liés à la tendance centrale de la distribution.

Paramètres de dispersion : ils montrent la manière dont les observations fluctuent autour de la tendance centrale.

III-1/ PARAMETRES DE POSITION

Calcul statistique du mode

Le mode d'un ensemble d'observations est la valeur la plus fréquemment rencontrée.

C'est la valeur la plus « à la mode ».

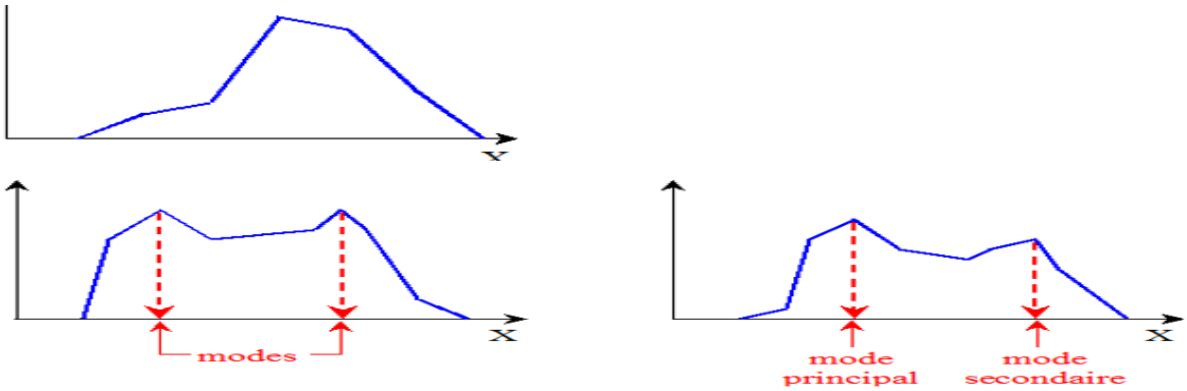
Le **mode** est un complément à la moyenne et à la médiane ; il permet de donner un indicateur de tendance centrale à un ensemble de données.

Dans le cadre de distributions regroupées en classe, le calcul du mode se fait à partir de la classe où la fréquence est la plus élevée avec la formule ci-dessous :

$$Mode = L + \left(\frac{d1}{d1+d2}\right) \times l$$

L = Limite inférieure de la classe identifiée la plus fréquentable, dite modale.
d1 = Différence entre la fréquence de la classe modale et la classe précédente.
d2 = Différence entre la fréquence de la classe modale et la classe suivante.
l = Longueur de la classe modale.

On appelle *distribution unimodale*, une distribution présentant un seul mode.



Une distribution bimodale est une distribution à deux modes.
Une *distribution multimodale* est une distribution présentant plusieurs modes. Elle est souvent le reflet d'une population composée de plusieurs sous-populations distinctes.

CALCUL DE MOYENNES

La **moyenne** d'un ensemble de données est une mesure de tendance centrale que suivent ces données. En statistique, la moyenne est aussi nommée espérance mathématique.

Il existe plusieurs types de moyenne qu'il est important d'appréhender selon les situations rencontrées : la moyenne arithmétique (plus courant), la moyenne arithmétique pondérée. On peut aussi citer la moyenne géométrique et la moyenne harmonique

LA MOYENNE ARITHMETIQUE

La moyenne arithmétique une méthode employée pour caractériser un ensemble de données et indiquer une tendance centrale.
Elle est la somme des observations divisée par le nombre n d'observations :

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Lorsque les données sont ordonnées sous forme de distribution de fréquence, la formule devient

$$\bar{x} = \frac{\sum_{i=1}^{p} x_i \cdot f}{\sum_{i=1}^{p} f_i}$$

Exemple 5 :

- Moyenne arithmétique classique dans une classe, la répartition des notes à un contrôle sont : 4, 5, 4, 8, 10, 7, 9, 6, 5, 2.

  La somme de ces notes : 4+5+4+8+10+7+9+6+5+2 = 60

  Sur 10 observations, la moyenne est donc 60 / 10 = 6.

- Moyenne arithmétique dans le cadre de fréquence :

Classe	Milieu de la classe	Fréquence	xi.Fi
0-50	25	3	75
50-100	75	4	300
100-150	125	2	250
150-200	175	6	1050
Somme totale		15	1675

La moyenne arithmétique est donc égale à : 1675 / 15 = 111,67.

LA MOYENNE ARITHMETIQUE PONDEREE

Ce type de calcul de moyenne est employé lorsque les observations n'ont pas toutes une importance identique. On attribue un poids à chaque observation dans le but de réaliser la pondération.

La moyenne arithmétique pondérée est donc égale à la moyenne des observations multipliées par leur poids, divisée par la somme des poids.

Exemple simple 6 :

1 étudiant obtient les notes de 12 en langage de programmation (coefficient 3), 8 en réseaux haut débits (coefficient 5) et 10 en analyse de trafic (coefficient 4).

Sa moyenne totale est donc égale à : (12×3 + 8×5 + 10×4) / 12 = 9,67.

MEDIANE

Valeur de la variable telle **qu'une moitié des valeurs lui soit supérieure ou égale et l'autre moitié des valeurs lui soit inférieure** ou égale est la médiane.
La **médiane** est donc la mesure se situant au centre d'un ensemble d'observations. Ces observations doivent être rangées par ordre croissant ou décroissant de façon à avoir 50% des observations de part et d'autre de la médiane.
Deux cas apparaissent suivant la parité de n.
Si nous avons un nombre d'observations impair, la médiane est donc la valeur du milieu.

n pair : $\boxed{\text{médiane} = \frac{1}{2} \cdot \left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right)}$

Si nous avons un nombre pair d'observations, la médiane sera la moyenne des valeurs des 2 observations du centre.

n impair : $\boxed{\text{médiane} = x_{\frac{n+1}{2}}}$ avec n : effectif total

Exemple 7 :

Soit un échantillon de 9 personnes dont le poids en kg est : 45 ; 68 ; 52 ; 56 ;62 ;63 ;68 ;74 ;89

Classés par ordre croissant :

45 – 49 – 52 – 56 – 62 – 63 – 68 – 74 – 89 kg

4 médiane 4

Si le nombre d'individus est pair, on prend la moyenne entre les deux valeurs centrales :

Par exemple si on a les poids 55, 45 ; 68 ;52 ;56 ;62 ;63 ;68 ;74 ;89 . En classant par ordre croissant, on obtient :

$$\underbrace{45 - 49 - 52 - 55 - 56}_{5} - \underbrace{62 - 63 - 68 - 74 - 89}_{5}$$

$$\text{médiane} = \frac{56 + 62}{2} = 59 \quad \text{kg}$$

En règle générale, si n est le nombre d'individus dans l'échantillon, la médiane porte le numéro d'ordre $\frac{n+1}{2}$ dans la suite des individus classés par ordre croissant

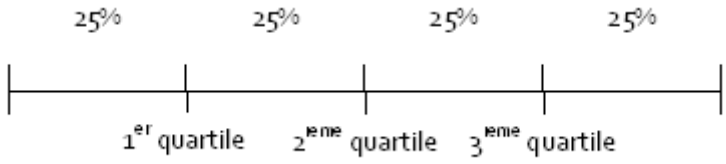
QUARTILES

Il s'agit des valeurs Q1, Q2 et Q3 de la grandeur mesurée qui partagent la série statistique de données en 4 parties d'effectifs à peu près identiques. Il est important de remarquer que le deuxième quartile Q2 est la médiane.

.

On calcul **(n+1)/4** pour le rang de Q1 et 3*(n+1)/4 pour le rang de Q3 ; si ces grandeurs ne sont pas des entiers, les quartiles ne sont donc pas des valeurs de la distribution. On réalise alors une interpolation.

Les quartiles sont des indicateurs de mesure de positions au sein d'une chaine d'observations. Les quartiles sont les quantiles qui divisent une distribution d'observations en 4 parties. Pour une distribution d'observations, nous cherchons donc 3 quartiles :



On généralise l'idée de médiane, en utilisant la notion de quantile où l'on va partager la distribution en n sous distributions de même taille.

Les valeurs de n classiques sont:

- n=4 --> quartiles
- n=10 --> déciles
- n=100 --> centiles

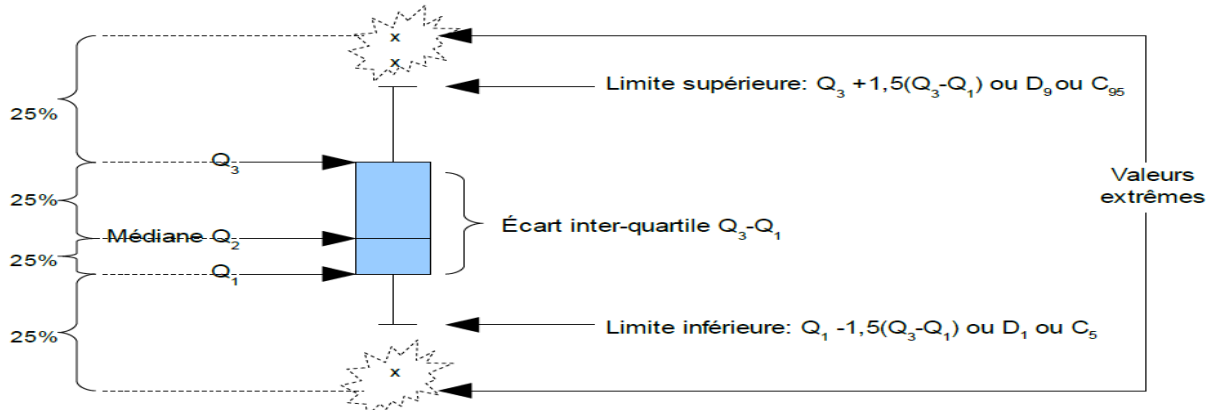
Ainsi la médiane est le deuxième quartile, le cinquième décile, le 50-ième centile (ou percentile)

50%), le quantile 0,5 ; Le premier quartile est le percentile 25%, le troisième quartile est le percentile 75%.En pratique, on peut procéder de la façon suivante :

- 25 % des valeurs observées de la série statistique soit inférieure à cette valeur.

Le troisième quartile est la première observation de la série statistique ordonnée telle que 75 % des valeurs observées de la série statistique soit inférieure à cette valeur

Application : box-plot ou boîte à moustache ou diagramme de Tuckey



ETENDUE INTERQUARTILE :

Il s'agit de l'intervalle contenant la moitié de la population autour de la médiane c'est à dire Q3 – Q1.

ETENDUE R :

L'étendue d'une distribution de données est la valeur obtenue en procédant à la différence entre la valeur minimale et la valeur maximale.

$$R = X_{max} - X_{min}$$

Exemple 8 :

On fait une étude statistique sur 10 sites de commerce électronique, ayant pour but de sonder sur une semaine le nombre de visiteurs et le nombre de commandes. On obtient le tableau suivant :

Le numéro du site (i)	1	2	3	4	5	6	7	8	9	10	
Le nombre de connexion (xi)	80	100	115	110	70	125	105	90	110	95	
Le nombre de commandes (yi)	32	50	62	56	8	80	62	50	62	38	

1. Calculer les moyennes arithmétiques de la variable statistique X et de la variable statistique Y.
2- calcul la médiane de X et de Y et comparer chacune des valeurs de médiane par rapport à la moyenne, pour X et pour Y.

III- 2 PARAMETRES DE DISPERSION

L'écart-type

L'écart type est la racine carrée de la variance.

Cet indicateur, fréquemment utilisé, permet de décrire la variabilité des valeurs d'un ensemble de données.

L'écart type est généralement utilisé pour compléter des indicateurs de tendance centrale tels la moyenne ou la médiane.

Plus la dispersion sera grande, plus l'écart type sera grand.

Généralement, l'écart type sera noté σ dès lors qu'il représente une population et S pour un échantillon. S est donc une estimation de σ.

Pour estimer l'écart type σ d'une population à l'aide d'un échantillon, nous pouvons calculer l'estimateur S à l'aide de la formule mathématique suivante :

$$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$

Pour obtenir un estimateur sans biais, il suffit de diviser non pas par n mais par **(n – 1)** (C'est pour estimer précisément la dispersion d'une population à partir d'un échantillon).

On obtient alors l'écart type

$$\sigma = \sqrt{\frac{1}{n-1}\sum\left(x-\overline{X}\right)^2}$$

Si l'écart type de la grandeur analysée dans la population n'est pas connu, on peut le remplacer par l'écart type calculé dans l'échantillon, pour autant que cet échantillon soit suffisamment grand.

$$\sigma\left(\overline{X}\right) \cong \frac{s}{\sqrt{n}} \qquad (si \quad n \geq 100)$$

PRECISION :

Notre échantillon est représentatif de la population, donc la moyenne sur **l'échantillon est donc une estimation de la moyenne sur la population.**
Nous désirons savoir quelle est la précision de cette estimation, afin de connaître de quelle quantité la vraie valeur est susceptible de s'écarter de notre estimation.
En fait, la précision va dépendre :

- de la taille de l'échantillon
- de la dispersion de la population

Dans une population peu dispersée, toutes les valeurs de l'échantillon seront forcément proches de la moyenne.
Dans une population plus dispersée, les valeurs de l'échantillon seront généralement plus éloignées de la moyenne. La moyenne de l'échantillon pourra donc s'écarter plus fortement de celle de la population.
Soient:

n le nombre d'individus dans l'échantillon,
σ l'écart type de la population

Alors, la précision de la moyenne peut être mesurée par un écart type sur la moyenne :

$$\sigma\left(\overline{X}\right)=\frac{\sigma}{\sqrt{n}}$$

La **précision sur la valeur moyenne sera donc d'autant meilleure que :**

1. **la population sera peu dispersée (σ petit)**
2. **l'échantillon sera grand (n grand)**

L'écart quadratique moyen (EQM)

Pour des raisons mathématiques, il est préférable, pour éliminer les signes (−), de calculer le carré des écarts plutôt que leur valeur absolue
On calcule donc la moyenne des carrés des écarts, puis on prend la racine carrée :

$$EQM=\sqrt{\frac{1}{n}\sum\left(x-\overline{X}\right)^2}$$

LA VARIANCE

la variance est définie comme la somme des différences au carré de chaque observation par rapport à la moyenne (déviation), divisée par le nombre d'observations.
En référence avec l'écart type, la variance sera notée S^2 lorsqu'il s'agit d'un échantillon et σ lorsqu'il s'agit d'une population.

COEFFICIENT DE VARIATION CV

Il représente une sorte d'écart-type relatif pour comparer les dispersions indépendamment des valeurs de la variable. Il s'exprime souvent en pourcentage.

$$CV=\frac{\text{écart type}}{\text{moyenne}}$$

Le coefficient de variation permet de comparer notamment la précision de Différents dosage, mesures, effectués avec le même appareil.
Par exemple :

a- calculer le coefficient de variation d'une distribution de moyenne 38,8 et d'écart type 3,13.

b- calculer le coefficient de variation d'une distribution de moyenne 95,9 et d'écart type 5,28.

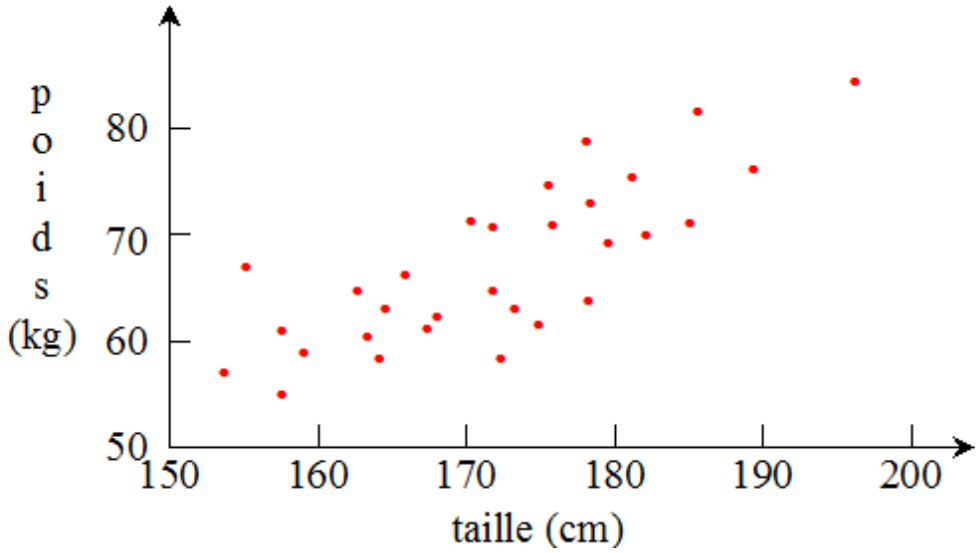
IV- REPRESENTATION DES DONNES

La première phase de l'analyse est la collecte des données **(phase Collecte des données). Le choix des classes**, soit leur nombre et leurs largeurs, n'est pas univoque. Il convient pour les déterminer de prendre en compte à la fois la nature de la distribution et le nombre de points de données. Souvent, dans le cadre d'une analyse de ce type, on utilise des classes de largeur identique.

Ensuite, vient la phase de traitement et d'analyse de données (calcul, représentation, inférence, prédiction, loi d'évolution, etc).

a- REPRESENTATION SOUS FORME DE NUAGE DE POINTS

Il s'agit des représentations des données sur un axe ou un repère. Si les données sont sous forme de couple, triplet etc, leur representation se fait sur un repère dimensionnel, tridimensionnel.



b- REPRESENTATION SOUS FORME DE DIAGRAMME CIRCULAIRE (SECTEUR)

Le degré d'un secteur est déterminé à l'aide de la règle de trois de la manière suivante : N $\longrightarrow$ 360◦ ni $\longrightarrow$ di (degré de la modalité i).

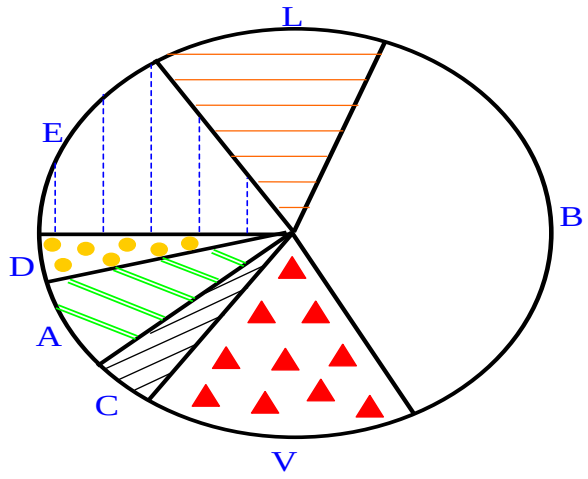
$$d_i = \frac{n_i \times 360}{N}.$$

Exemple 9 :

On donne le nombre d'enfants ni par famille xi.

xi	0	1	2	3	4	5	6
ni	18	32	66	41	32	9	2

Représenter le diagramme en pie des données



c- REPRESENTATION SOUS FORME D'HISTOGRAMME

Pour un nombre élevé de mesures, d'observation, ou s'il y a plusieurs valeurs identiques, on regroupe parfois les valeurs de la série statistique étudiée en classes. Une distribution de fréquences est un tableau des fréquences associées à ces classes.

Un histogramme est une représentation graphique dans laquelle les rectangles représentés ont des largeurs proportionnelles aux amplitudes des classes et des aires proportionnelles aux fréquences de ces classes. L'histogramme permet de visualiser rapidement des données.

Pour une représentation correcte d'un histogramme, il faut surtout éviter d'utiliser un nombre de classes mal adapté. Des règles régissent la construction des Histogrammes.

- **nombre de classes k :** $k \approx \sqrt{n}$ avec n le nombre total de mesures.

Généralement on ne dépasse pas 20 classes.

La formule de Sturge

$$k = 1 + 3.3 \log_{10}(N).$$

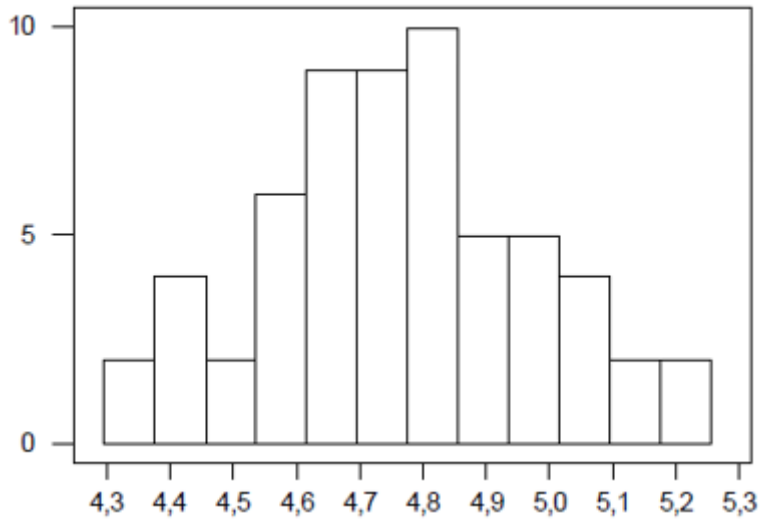
La formule de Yule

$$k = 2.5 \sqrt[4]{N}.$$

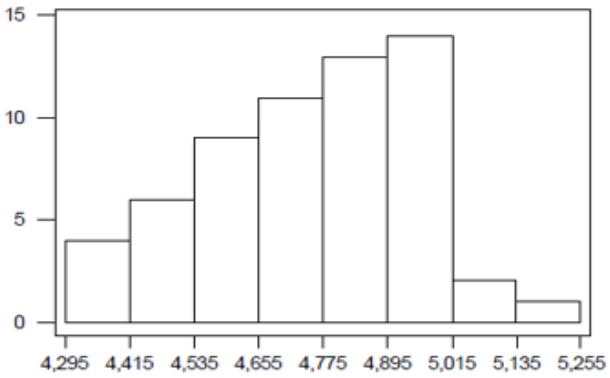
- **intervalle de classe h :** $h \approx \frac{R}{k}$ arrondi au multiple immédiatement supérieur de la résolution de mesure où R est l'étendue de mesure.

- **limite inférieure de la première classe :** x_{min} – ½ résolution de mesure

Histogramme bimodal : il révèle une population hétérogène composé d'un mélange de plusieurs populations



Histogramme tronqué aux extrémités : il peut révéler une suppression d'un certain nombre de valeurs expérimentales extrêmes sur un document de fabrication.



HISTOGRAMMES NORMALISES ET CALCUL DE PROBABILITE

Approche par les exercices :

Exercice 1 : On donne les valeurs de marge de puissance en dB d'un Canal :

4,62 4,45 5,08 4,83 4,74 4,68 4,74 4,59 4,77 4,67 5,08 4,79 5,03 5,20 4,74 4,96 4,47 4,60 4,72 4,80
4,43 4,81 4,73 4,83 4,96 4,97 4,88 4,68 4,75 4,79 4,57 4,98 4,84 4,43 4,35 5,11 4,68 4,30 5,20 4,92
4,55 4,91 5,17 4,61 4,56 5,00 4,63 4,71 4,84 5,07 4,88 4,70 4,63 4,85 4,90 4,53 4,67 4,79 4,69 4,43
La précision de la mesure est de 0,12.

1- Calculer L'intervalle de classe h.
2- Donner la limite inferieures de la première classe
3- Donner le tableau des classes et des effectifs.
4- Représenter les histogrammes sous matlab
5- Donner la moyenne, l'écart type et la variance sous matlab
6- Ecrire un script Matlab qui répond dans un seul programme aux 5 questions ci-dessus. Commenter

Exercice 2 : On donne les valeurs de marge de puissance en dB d'un Canal de transmission par le vecteur suivant :

```
data = [0.0651 0.0548 0.0461 0.0686 0.1268 0.2266
0.2292 0.1187 0.0299 0.0146 0.0092 0.0048 0.0032
0.0024 0.0470 0.0594 0.0743 0.0918 0.1120 0.1352
0.1611 0.1897 0.2208 0.2538 0.2884 0.3238 0.3592
0.3937];
```

La précision de la mesure est de .012

1- Calculer L'intervalle de classe h.
2- Donner la limite inferieures de la première classe
3- Donner le tableau des classes et des effectifs.
4- Représenter les histogrammes sous Matlab
5- Donner la moyenne, l'écart type et les variances sous Matlab
6- Ecrire un script Matlab qui répond dans un seul programme aux 5 questions ci-dessus. Commenter

- **Exercice 3** : On donne le script suivant en Matlab ou x désigne l'âge et y les pulsions cardiaques d'une population cible donnée.
-
```
x=20:5:70
y=[150 146 142 139 135 131 127 124 120 116 112]
plot(x,y,'*') % trace de la courbe age en function des pulsions cardiaques
a1=polyfit(x,y,1)
% a1= -0.7527  164.9636
% on trouve un polynome qui decrit la variation de x en fontion de y c'est un polynome de degré 1. Interpolation lineaire par cepolynone de degre 1
hold on
xi=20 :0.5 :70
yi=polyval(a1,xi);
plot(xi,yi,'c')
x1= linspace(20,0.1,70)
y2=polyval(a1,x);
erro= ((y-y2)./y).*100 ;
mean(abs(erro)) ;
%disp(`[les pourcentage d'erreur est de :',num2str(mean(abs(erro))])
disp(['les pourcentage d'erreur est de  vaut ', num2str(mean(abs(erro)))])
xlabel('Age (années)')
ylabel('pulsions cardiaque')
```
- Commenter toutes lignes du programme qui ne le sont pas.
- Faire un interpolation avec un polynôme de degré 2. Quelle est la meilleure entre l'interpolation avec un polygone de degrés 1 et 2 dans notre cas ? expliquer
- Retrouver les paramètres de la fonction d'interpolation de degré 1 en calculant les différents coefficients (y= ax+b). comparer avec ce qui est trouvé avec Matlab

Exercice 4 : On cherche à étudier la relation entre le nombre d'enfants d'un couple et son salaire. On dispose de la série bidimensionnelle suivante :

Salaire en KFCF(Y)	Nombre d'enfants (X)
510	4
590	3
900	2
1420	1
2000	0
600	5
850	6
1300	7
2200	8

a) Calculer le coefficient de corrélation linéaire entre ces deux variables statistiques. Conclusion ?
b) Un expert en démographie affirme que les deux caractéristiques sont indépendantes. Qu'en pensez-vous ?