Introduction à l'Analyse de données

I- Introduction

Définition :

Les données sont des informations issues d'observations, de mesures faites sur une population humaine, animale ou de chose/de choses (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 (ou unités statistiques) auxquels on décide de sintériser. Sa taille, est généralement notée par **N**, elle peut est grande, ou même infinie.

Variable (ou caractère) statistique, valeurs : Une variable est une information dont on recueille (ou observe ou mesure) la **valeur** 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. C'est à partir des valeurs observées que le statisticien ou l'analyste de données rétablit ses classements d'individus.

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

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

Recensement : Recueil des valeurs de la totalité des individus de la population. Les valeurs recueillies sont les données.

Sondage, *n*-échantillon, base de sondage, taux de sondage : Un sondage est le recueil des valeurs d'une partie (l'**échantillon**) d'effectif *n* (d'où l'expression *n*-échantillon) de la population (dite ***base de sondage***). Le taux de sondage est le rapport *n*/**N**.

Variable, (ou caractère) qualitatif (ou nominal) : 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".

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 consommateurs.

Variable, ou caractère quantitatif : Variable *numérique* telle qu'on peut calculer, par une opération mathématique quelconque, comme l'addition, pour deux (ou plusieurs) individus, une valeur appelée **total**, à partir des valeurs de ces individus. Pour l'addition il s'agit de la **somme**. Exemple des ventes annuelles = somme des ventes de l'ensemble de tous les jours ouvrés de l'année. Les valeurs observées forment un ensemble continu ou non, infini ou non.

Exemple1 :

Sur chacun des individus sondés, on observe un **caractère** (ou **variable**). Par exemple :

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

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

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

- métier
- fumeur
- titulaire du permis B

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

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

Exemple 3 :- 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 exemple 3 :

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,	Petite, moyenne, grande, faible	Acceptable, bonne, excellente, mauvaise, insuffisante
Affaiblissement en dB de signal,	Petite, moyenne, grande, faible, nulle	Acceptable, bonne, excellente, mauvaise, insuffisante
Rapport signal sur bruit.	Petite, moyenne, grande, faible, elevée	Acceptable, bonne, excellente, mauvaise, insuffisante

II- ECHANTILLONNAGE :

1- Echantillonnage aléatoire simple

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

Pour obtenir un échantillon de cette sorte, on numérote 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.

2- Echantillonnage par grappes

Pour l'**échantillonnage par grappes**, il faut diviser la population en grappes, c'est-à-dire en sous-ensembles de façon à ce que chacun de ces sous-ensembles devant être représentatif de la population mère.

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

Exemple 4: Des études menées à l'échelle d'une ville, que le fait que l'on divise en quartiers constitue un exemple d'échantillonnage par grappes.

Pour obtenir un échantillonnage par grappes ayant les propriétés statistiques aussi précises que possible, il faut :

- Un nombre de grappes non conséquent
- La taille des grappes uniformes
- Une homogénéité des individus composant les grappes

3- Echantillonnage par la méthode des quotas (utilisée en sondage)

La **méthode des quotas** est basée sur la répartition connue d'une population (âge, sexe, situation géographique, catégorie socio-professionnelle…).

Une fois la dimension et les critères du sondage que l'on souhaite, effectué, il suffira alors de calculer le nombre d'individus par chaque critère choisi.

Cependant, cette méthode (la moins onéreuse) a des limites qu'il faut préciser et qui permettent de comprendre pourquoi les sondages lus régulièrement apportent plus des tendances de l'opinion plutôt que de chiffres véritablement précis :

- Cette méthode repose sur l'hypothèse que l'information que l'on souhaite obtenir est corrélée avec la population. Ce n'est qu'une hypothèse de représentativité qui est difficile à démontrer voire impossible.
- Le choix des individus sélectionnés par des enquêteurs lors de la méthode des quotas ne permet pas de calculer des probabilités d'appartenance à l'échantillon. Ceci entraine une difficulté de calcul d'erreurs et donc de précision de l'analyse.

- **Les quotas et l'aspect mathématique**

Si l'on part d'une population telle que décrite dans le tableau ci-dessous :

Sexe		Age		Pays	
Masculin	600	Moins de 25 ans	200	France	250
Féminin	400	Entre 25 et 60 ans	500	Espagne	200
		Plus de 60 ans	300	Italie	250
				Suisse	150
				Belgique	150
Total	1000	Total	1000	Total	1000

Et que nous décidons un taux de sondage de 1/5 (20%), nous interrogerons donc 200 personnes avec la répartition suivante :

Sexe		Age		Pays	
Masculin	120	Moins de 25 ans	40	France	50,00
Féminin	80	Entre 25 et 60 ans	100	Espagne	40,00
		Plus de 60 ans	60	Italie	50,00
				Suisse	30,00
				Belgique	30,00
Total	200	Total	200	Total	200

Le choix des individus au sein de ces échantillons se réalise de manière aléatoire. Bien évidemment un petit tableau croisé dynamique permettant de représenter la juste distribution des segments est nécessaire avant de pratiquer le tirage.

La méthode de la boule de neige

La méthode de la boule de neige consiste à diffuser un questionnaire à des personnes connues ayant les caractéristiques recherchées puis de leur demander d'indiquer d'autres personnes de profil similaire. L'échantillon n'est pas représentatif mais la méthode est simple et adaptée au lancement d'une nouvelle activité.

Echantillonnage et questionnaire

On procède de façon implicite en élaborant un questionnaire en ligne et en recueillant les réponses et en les analysants. Il existe plusieurs sites d'élaboration de questionnaire de sondage et d'enquête en lignes dont :

https://www.sli.do/

https://fr.surveymonkey.com/

https://www.google.com/intl/fr-CA/forms/about/

Les réponses aux questions peuvent être traitées par : tri à plat ou par tri croisé.

Le tri à plat, qui donne la répartition des réponses question par question, est le premier traitement statistique effectué : il permet d'avoir une première idée des résultats et constitue naturellement la base des rapports d'enquête.

Le tri à plat est le premier des traitements statistiques d'une enquête ou sondage. Le tri à plat vous apporte une première connaissance des données recueillies grâce à votre questionnaire en ligne et il va vous permettre d'aller plus loin dans votre analyse.

Le tri croisé est un traitement des résultats d'une enquête, portant généralement sur deux questions, et consistant à indiquer dans un tableau la répartition des différentes combinaisons de réponses.
Dans le traitement des résultats d'une enquête ou sondage, les tris croisés consistent à mettre en relation les réponses à des questions différentes pour rechercher quels critères jouent les uns sur les autres. Parfois complexes à interpréter, ils sont essentiels pour affiner l'analyse des résultats.

Remarque : Minimiser dans le traitement des questions/questionnaires les erreurs de mesure et de

Minimiser les erreurs de mesure
L'inexactitude ou l'imprécision des réponses, appelée erreur de mesure, peut parfois être imputée à la mauvaise foi des répondants.

Erreur de couverture
L'erreur de couverture est souvent le principal problème évoqué à propos des sondages en ligne.

Les erreurs et le biais joue sur la fiabilité de l'enquête ou du sondage

III- PARAMETRES 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. On distingue : La **moyenne**, la mode, **la médiane**, **les Quartiles**.
Paramètres de dispersion : ils montrent la manière dont les observations

Fluctuent autour de la tendance centrale. On distingue : **L'écart-type, La variance, le Coefficient de variation**

1/ Paramètres de position

a- Calcul de moyennes en statistiques

La **moyenne** d'une variable aléatoire est une mesure de tendance centrale de cette variable. C'est l'indicateur le plus utilisé arithmétique. 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
- La moyenne arithmétique pondérée
- La moyenne géométrique
- La moyenne harmonique

La moyenne arithmétique

La moyenne arithmétique utilisée pour caractériser un ensemble de données et indiquer une tendance centrale.

La moyenne arithmétique est la somme des observations divisée par le nombre n d'observations :

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

Lorsque les données sont ordonnées sous forme de distribution de fréquence, la formule de la moyenne est donnée par :

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

Exemples de calcul 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 :
-

Classes	Milieux de la classe Xi	Fréquences Fi	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 arithmétique pondérée

Ce type de calcul de moyenne est employé lorsque les observations n'ont pas toutes une importance identique.

Il est donc attribué un poids à chaque observation afin 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.

L'exemple simple et parlant est celui de l'examen 6:

1 élève obtient les notes de 14 en français (coefficient 3), 16 en mathématique (coefficient 5) et 17 en anglais (coefficient 4).

Sa moyenne totale est donc égale à : (14×3 + 16×5 + 17×4) / 12 = 15,83.

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. Calculer les écarts-type de la variable statistique X et de la variable statistique Y.
3. Calculer la covariance entre X et Y.
4. Calculer le coefficient de corrélation linéaire entre X et Y. Commenter. 5. Déterminer la droite de corrélation Y = aX + b.

b- Calcul statistique du mode

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

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

Dans le cadre de distributions regroupées en classe, le calcul du mode se fait avec la formule ci-dessous :

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

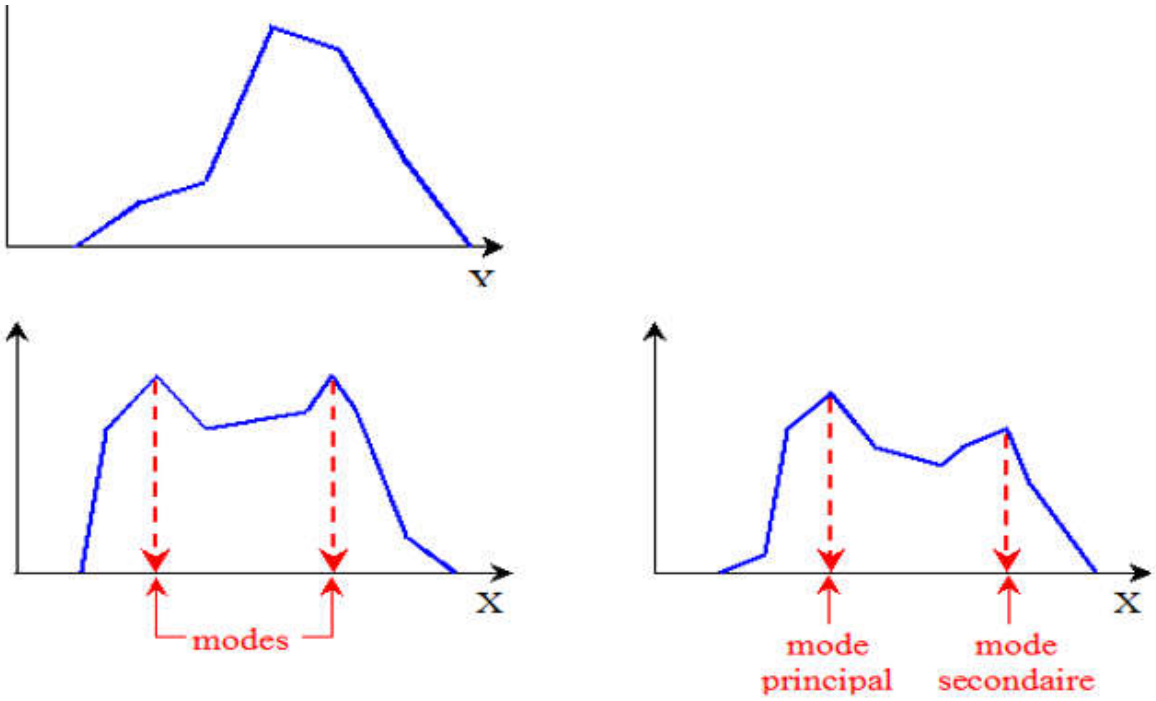
L = Limite inférieur de la classe identifiée la plus fréquentable, dite modale.
d1 = Différence entre la fréquence de la classe modale et la classe précédente.
d2 = Différence entre la fréquence de la classe modale et la classe suivante.
l = Longueur de la classe modale.

Il correspond au sommet de la distribution:

le mode est la valeur la plus fréquente

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

On appelle distribution unimodale, une distribution présentant un seul mode



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

c- La Médiane :

C'est la 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. Deux cas apparaissent suivant la parité de n.

Si nous avons un nombre d'observations impair, **la médiane est donc la valeur du milieu.**

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

Il est souvent intéressant de lier le calcul de la moyenne avec la médiane. Parce que si les observations listées contiennent de nombreuses données extrêmes, la médiane permet un éclairage de la mesure centrale.

la médiane est la valeur pour laquelle il y a autant d'individus à gauche qu'à droite dans l'échantillon

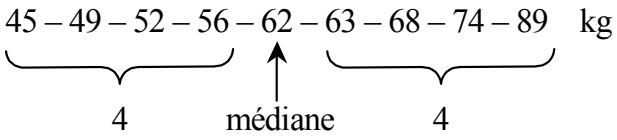
Pour déterminer la médiane d'un échantillon ou d'une population :
(1) on classe les individus par ordre croissant
(2) on prend celui du milieu

Exemple 7:

- Soit un échantillon de 9 personnes dont le poids est :

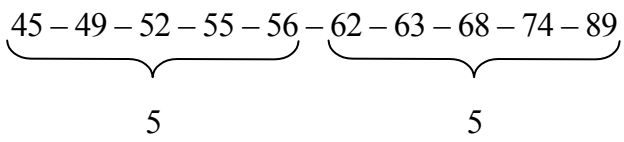
45 – 68 – 89 – 74 – 62 – 56 – 49 – 52 – 63 kg

classés par ordre croissant : 45 ; 68 ;52 ;56 ;62 ;63 ;68 ;74 ;89



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

45 ; 68 ;52 ;56 ;62 ;63 ;68 ;74 ;89



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

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

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

moyenne : n : effectif total $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

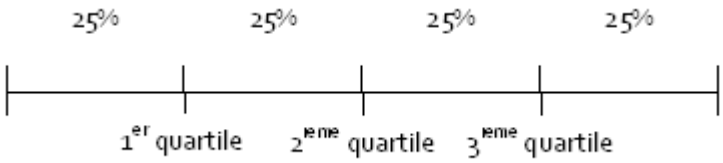
d- Les Quartiles :

Valeurs des quartiles Q1, Q2 et Q3 de la grandeur mesurée qui partagent la série statistique en 4 parties d'effectifs à peu près identiques.
Q2 est la médiane. Le calcul de Q1 et Q3 diffère légèrement suivant les auteurs ou les logiciels.

On calcule $\frac{n+1}{4}$ pour le rang de Q_1 et $3 \cdot \left(\frac{n+1}{4}\right)$ pour le rang de Q_3

Si ces grandeurs ne sont pas des entiers, les quartiles ne sont donc pas des valeurs de la distribution. On réalise alors une interpolation.

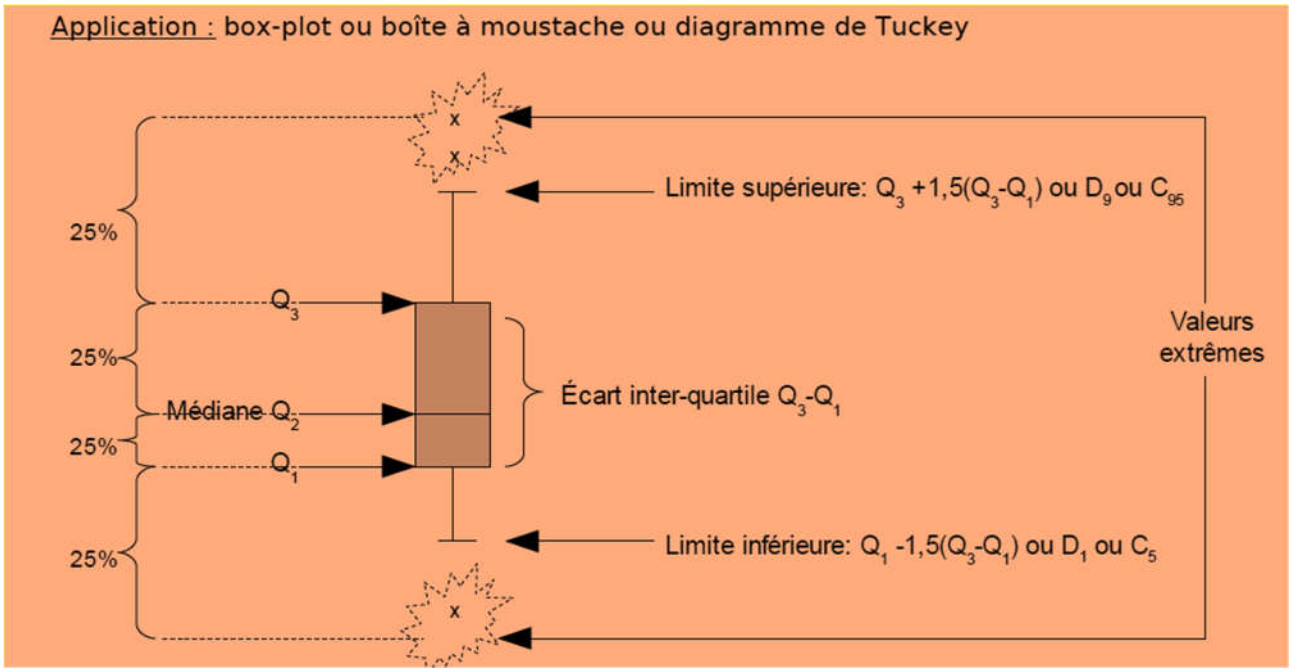
Les quartiles sont des indicateurs de mesure de positions au sein d'une chaine d'observations. Pour une distribution d'observations, nous cherchons donc 3 quartiles :



Pour préciser cette idée de répartition, on dispose d'autres outils:
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, pour la recherche des quartiles, on adopte la définition suivante :
Le premier quartile est la première observation de la série statistique ordonnée telle que
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



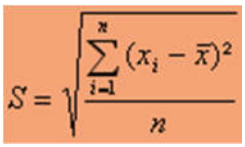
Étendue interquartile :
Intervalle contenant la moitié de la population autour de la médiane c'est à dire Q3 – Q1
Étendue R : c'est la différence entre la plus grande et la plus petite valeur d'une observation, d'une VA, d'une population, d'une distribution.

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

2/ Paramètres de dispersion

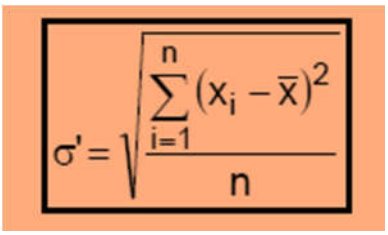
a- L'écart-type

L'écart type est une mesure de dispersion. Cet indicateur, communément employé, 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.
L'objectif est donc de voir si les valeurs d'un ensemble de données sont plus ou moins regroupées autour de la tendance centrale.
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 un échantillon, nous pouvons calculer l'estimateur S de l'écart type par :



Pour obtenir un estimateur sans biais, il suffit de diviser non pas par n mais par (n – 1).

On peut aussi écrire que :



L'écart-type est le plus utilisé des paramètres de dispersion.

L'étendue est beaucoup plus facile à calculer mais elle donne une valeur très Imprécise de la "largeur de la répartition" quand le nombre de valeurs est supérieur à 10.

b- La variance

La variance aussi une mesure de dispersion d'une distribution d'une variable aléatoire.

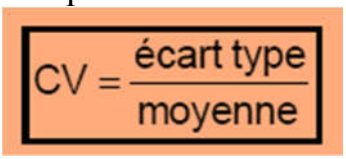
C'est la somme des différences au carré de chaque observation par rapport à la moyenne (déviation), divisée par le nombre d'observations.

En lien avec l'écart type, la variance sera notée S^2 lorsqu'il s'agit d'un échantillon et σ^2 lorsqu'il s'agit d'une population.

Variance : il s'agit de la moyenne des carrés des écarts à la moyenne.

c- 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.



Le coefficient de variation permet de comparer notamment la précision de différent 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.

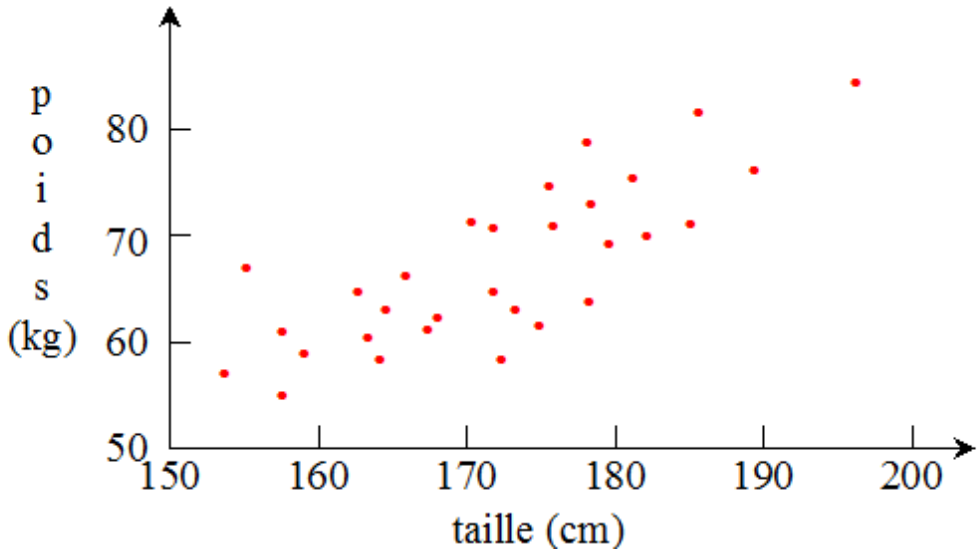
IV- REPRESENTATION DES DONNES

La première phase de l'analyse est la collecte des données **(phase Collecte des données)**

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

a- Représentation Sous Forme De Nuage De Points

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



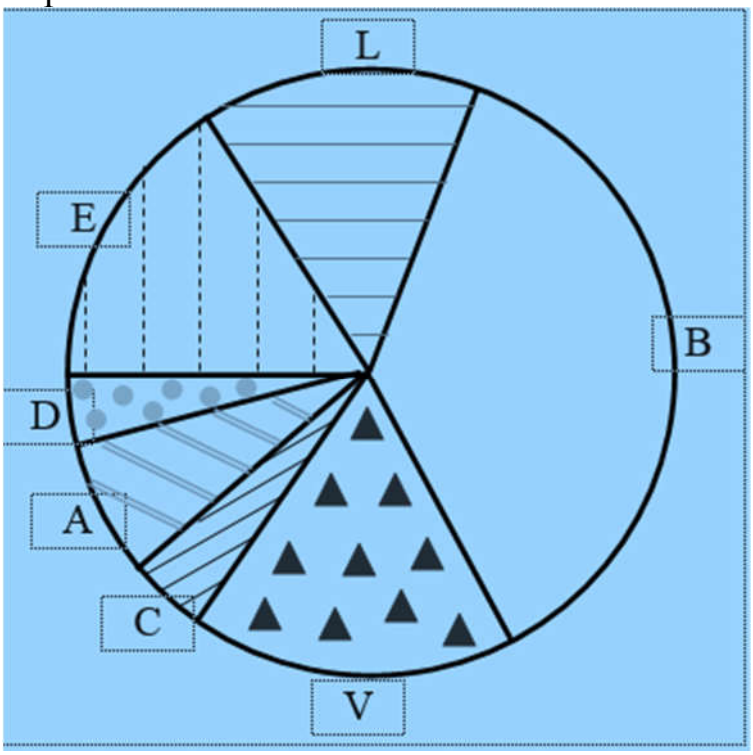
b- Représentation Sous Forme De Diagramme Circulaire (Secteur)

Le degré d'un secteur est déterminé à l'aide de la règle de trois de la manière suivante :

N → 360◦ ni → di (degré de la modalité i).

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

Exemple 8 : On dispose de 10 billes rouges(V), 42 billes bleues (B), de 11 billes mauves (L), de 14 billes bleu –clair (E), de 7 billes jaunes (D), de 9 billes vert clair(A) et de 7 billes grises (C)
Représenter les données sous forme circulaire.



c- Représentation 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.

- **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

2. Une réponse : la formule de Sturge

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

3. Une réponse : la formule de Yule

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

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.
On pourra trouver dans la littérature de nombreuses suggestions de choix pour le nombre de classe. Citons par exemple :

Exemple : Soit la masse d'une préparation culinaire avant conditionnement. Le calcul d'amplitude de classe donne h_{th} = 0,014 kg. La résolution de la balance utilisée est de 0,001 kg. On arrondit la valeur h à 0,015 kg.

Les classes peuvent être du type **[limite inférieure ; limite supérieure[ou] limite inférieure ; limite supérieure**

L'intervalle de classe h est: $h = \frac{0,90}{8} \approx 0,12$ *en arrondissant au multiple immédiatement supérieur de la résolution de mesure (0,01 ici).*

La limite inférieure de première classe est 4,30 – 0,5 . 0,01 = 4,295.

On en déduit alors le tableau suivant des classes et effectifs.

Intervalle de classe	*Effectif de la classe*
4,295 – 4,415	*2*
4,415 – 4,535	*6*
4,535 – 4,655	*9*
4,655 – 4,775	*15*
4,775 – 4,895	*12*
4,895 – 5,015	*8*
5,015 – 5,135	*5*
5,135 – 5,255	*3*

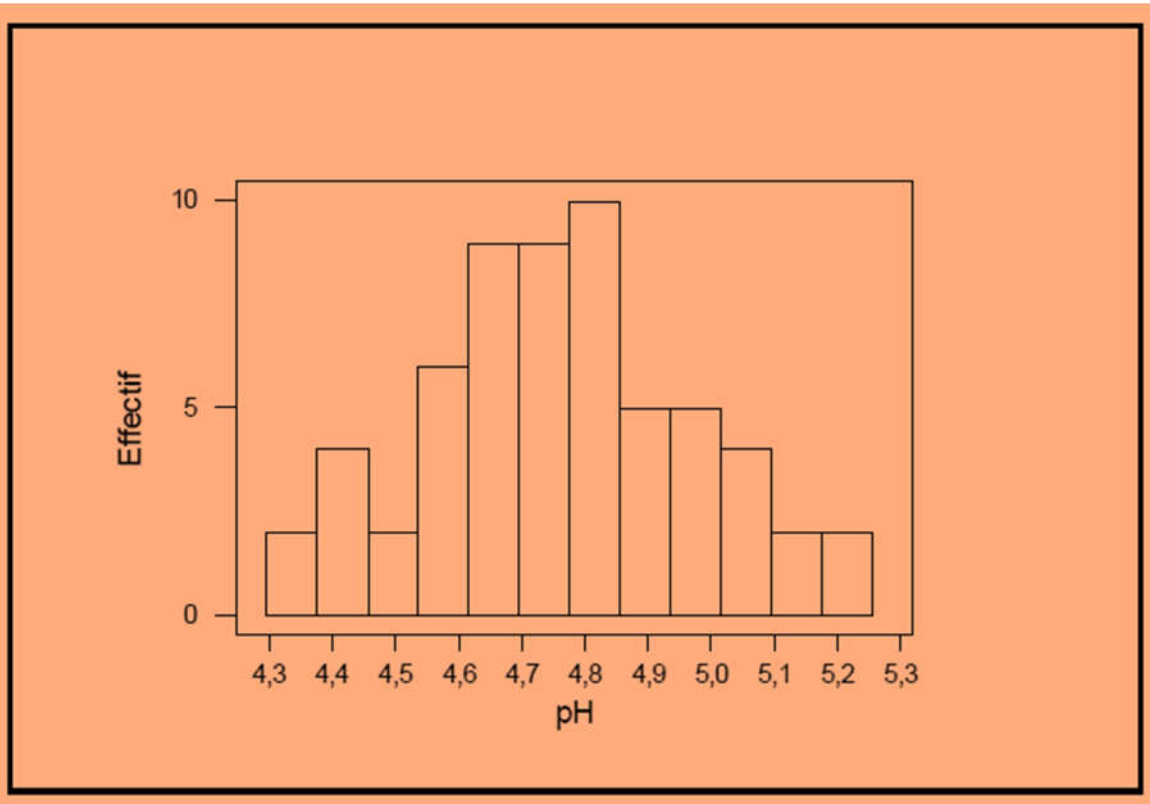
Histogramme bimodal :

Il révèle une population hétérogène composé d'un mélange de plusieurs populations. Ainsi on peut déceler dans une production livrée un mélange de 2 lots. Dans l'exemple précédent un facteur expérimental non contrôlé pouvant prendre deux valeurs (par exemple deux lots de matière première) entraîne deux "populations de pH".

0,547	0,563	0,532	0,521	0,514	0,547	0,578	0,532	0,552	0,526	0,534	0,560	0,502	0,503	0,516	0,565
0,532	0,574	0,521	0,523	0,542	0,539	0,543	0,548	0,565	0,569	0,574	0,596	0,547	0,578	0,532	0,552
0,554	0,596	0,529	0,555	0,559	0,503	0,499	0,526	0,551	0,589	0,588	0,568	0,564	0,568	0,556	0,523
0,526	0,579	0,551	0,584	0,551	0,512	0,536	0,567	0,512	0,553	0,534	0,559	0,498	0,567	0,589	0,579

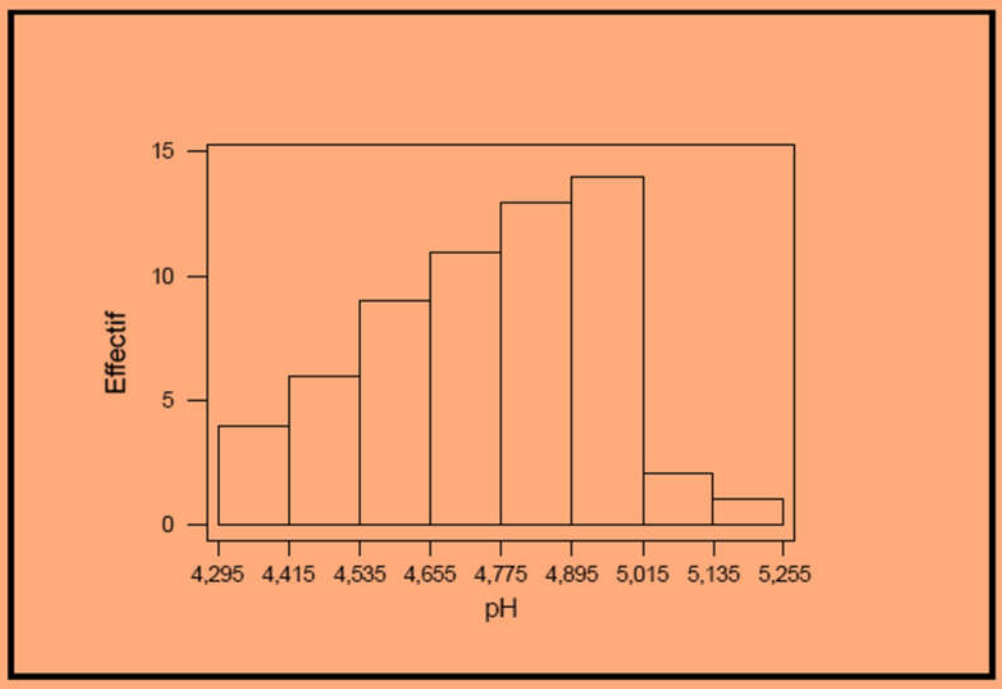
Histogramme bimodal :

Il révèle une population hétérogène composé d'un mélange de plusieurs populations



Les histogrammes permettent aussi de relever des anomalies:

• **histogramme tronqué aux extrémités :** il peut révéler une suppression d'un certain nombre de valeurs expérimentales extrêmes sur un document de fabrication … afin par exemple de ne pas "sortir" des limites de contrôles !



Exercice : 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

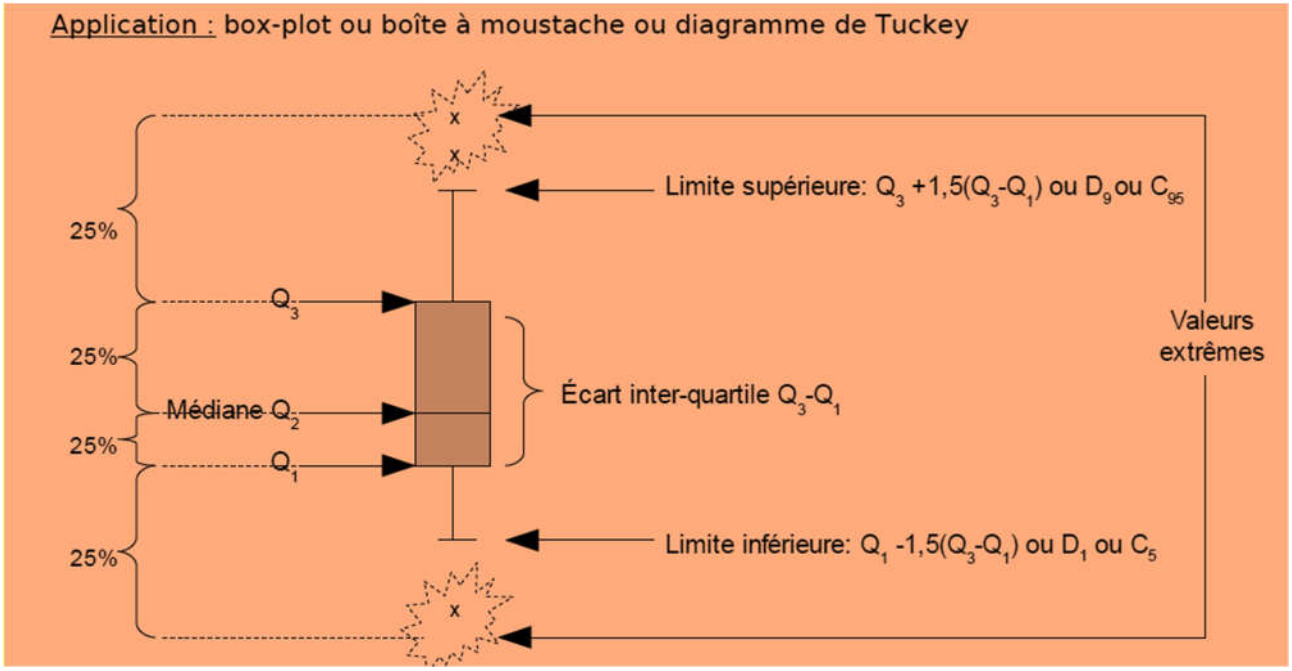
La précision de la mesure est de 0,12.

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

3/ Autre représentation des données : "la boîte à moustaches"

Les boîtes à moustaches permettent de comparer visuellement deux échantillons sur des critères de forme, de dispersion et de centrage des données.
La bordure inférieure de la boîte représente le premier quartile (Q1) et la bordure supérieure représente le troisième quartile (Q3). La portion du diagramme comprise dans la boîte représente donc l'étendue interquartile ou la moitié centrale (50 %) des observations.
La ligne horizontale qui traverse la boîte représente la médiane des Données. Les lignes qui sortent de la boîte sont appelées moustaches. Les moustaches s'étendent vers l'extérieur pour indiquer à, leurs extrémités la valeur la plus basse et

La valeur la plus haute dans la série (à l'exception des valeurs aberrantes). La boîte à moustaches permet aussi d'évaluer la symétrie des données :
Lorsque les données sont symétriques, la ligne médiane se situe à peu près au milieu de la boîte interquartile et les moustaches sont de la même longueur.
Si les données sont asymétriques, il se peut que la médiane ne tombe pas au milieu de la boîte interquartile et une moustache peut être nettement plus longue que l'autre.

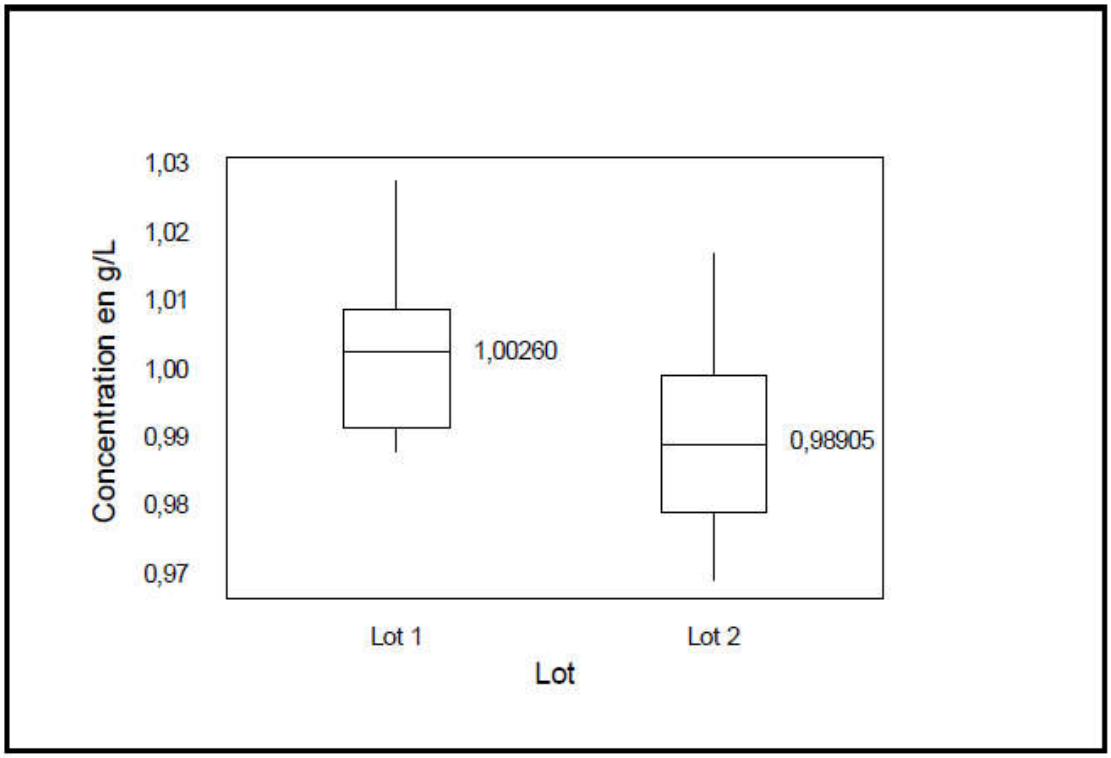


Exemple 9 ***:*** *on considère deux lots de solutions étalons en ions nickel II. On analyse par spectrophotométrie 16 flacons de chaque lot qu'on souhaite comparer. A partir des concentrations du tableau suivant, on réalise une boîte à moustaches pour chaque lot.*
Lot 1 :1,0047 1,0089 0,9922 0,9880 1,0120 0,9932 1,0005 1,0089 0,9911 1,0057 0,9901 1,0277 1,0120 0,9880 0,9995 1,0057

Lot 2 : 0,9807 0,9713 0,9984 0,9995 0,9786 0,9828 0,9692 0,9922 0,9838 0,9911 1,0099 1,0172 1,0068 0,9870 0,9723 0,9943

Représenter les donner suivant une boite à moustache.

Lot 1	*1,0047*	*1,0089*	*0,9922*	*0,9880*	*1,0120*	*0,9932*	*1,0005*	*1,0089*
	0,9911	*1,0057*	*0,9901*	*1,0277*	*1,0120*	*0,9880*	*0,9995*	*1,0057*
Lot 2	*0,9807*	*0,9713*	*0,9984*	*0,9995*	*0,9786*	*0,9828*	*0,9692*	*0,9922*
	0,9838	*0,9911*	*1,0099*	*1,0172*	*1,0068*	*0,9870*	*0,9723*	*0,9943*



Le lot 2 a une concentration moyenne plus faible que le lot 1; l'examen des boîtes interquartiles montre une répartition plus symétrique des données du lot 2, au moins dans la partie centrale de la distribution.

Le calcul des quartilesmontre les résultats suivants :

Lot	Q_1	Q_2	Q_3
Lot 1	0,9914	1,0026	1,0089
Lot 2	0,9791	0,9891	0,9992

Pour obtenir ces résultats il faut classer les 16 valeurs de chaque lot. Pour le lot 1, on obtient la médiane Q_2 avec la relation donnée plus haut :

$$Q_2 = \frac{1}{2}\cdot\left(x_{\frac{16}{2}} + x_{\frac{16}{2}+1}\right) = \frac{1}{2}\cdot(x_8 + x_9) = 0{,}5\cdot(1{,}0005 + 1{,}0047) = 1{,}0026$$

Pour le quartile Q_1, l'application de la relation $\frac{n+1}{4}$ pour déterminer le rang entraîne une valeur de rang de $\frac{16+1}{4} = 4{,}25$ qui n'est pas entière. Donc Q_1 s'obtient par interpolation entre les valeurs x_4 et x_5 soit 0,9911 et 0,9922.

Donc $Q_1 = 0{,}9911 + (4{,}25 - 4)\cdot(0{,}9922 - 0{,}9911) = 0{,}9914$.

Exercice 1 : 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 2: 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.

7- Calculer L'intervalle de classe h.
8- Donner la limite inferieures de la première classe
9- Donner le tableau des classes et des effectifs.

10- Représenter les histogrammes sous Matlab
11- Donner la moyenne, l'écart type et la variance sous Matlab
12- 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 âge en fonction des pulsions cardiaques
a1=polyfit(x,y,1)
% a1= -0.7527  164.9636
% on trouve un polynôme qui décrit la variation de x en fonction de y c'est
un polynôme de degré 1. Interpolation linéaire par ce polynôme de degré 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')
```

1- Commenter toutes lignes du programme qui ne le sont pas.
2- Faire une interpolation avec un polynôme de degré 2. Quelle est la meilleure entre l'interpolation avec un polynôme de degré 1 et 2 dans notre cas ? expliquer
3- 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.