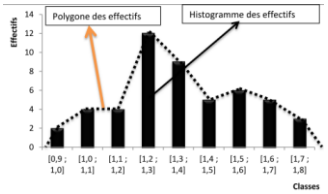
UNIVERSITE DE SARH **ANNEE ACADEMIQUE 2024/2025**

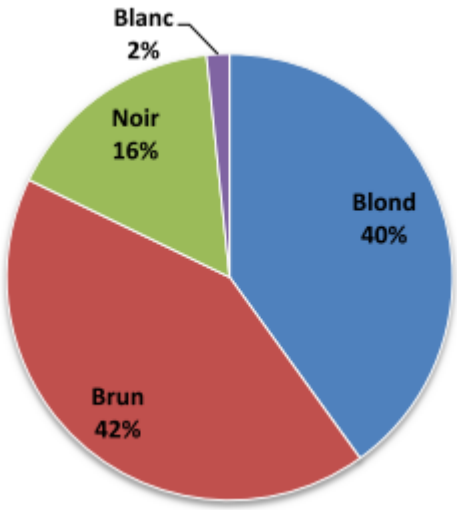
IUSAE-IUDRH/FLASH LICENCE 1

$$\begin{cases} Me = \frac{Xn+1}{2} & \text{Si n est impair} \\ Me = \frac{X\frac{n}{2}+X\frac{n}{2}+1}{2} & \text{Si n est pair} \end{cases}$$

$$Mo = e_{j-1} + a_j \times \frac{Dj-1}{Dj-1+Dj}$$



COURS DE STATISTIQUE DESCRIPTIVE



Chargé de cours : TAMTIAL NGARIBAN

Enseignant-Chercheur à l'Institut Universitaire des Sciences Agronomiques et de l'Environnement de l'Université de Sarh. Doctorant (2ème Année) en Agronomie/Cultures maraichères à l'Université de N'Djamena/Ecole Doctorale Sciences-Techniques-Environnement. Chef de Service de la Scolarité et des Statistiques à la FSAE/UDS

OBJECTIFS DU COURS

Le cours a pour but d'initier les étudiants aux principes de base de la statistique. Le cours vise principalement à introduire et faire méditer les concepts fondamentaux et méthodes élémentaires de la statistique pour permettre un apprentissage autonome ultérieur de méthodes complémentaires.

On veut développer le sens critique nécessaire lors de la mise en œuvre et de l'interprétation d'un traitement statistique. Pour cela, on introduira et utilisera un cadre mathématique rigoureux. Nous fournirons autant d'exemples et de figures nécessaires afin d'obtenir une meilleure compréhension du cours.

La statistique descriptive a pour but d'étudier un phénomène à partir des données. Cette description se fait à travers la présentation des données (la plus synthétique possible), leur représentation graphique et le calcul de résumés numériques.

HISTOIRE DE LA STATISTIQUE

L'histoire de la "statistique" remonte à une époque très ancienne. Les activités statistiques (dénombrements) ont commencé bien avant la création du mot, l'application de la méthode et de l'analyse statistique.

Depuis l'antiquité, les Empereurs, les Rois et les Hommes d'Eglise réalisaient des dénombrements de populations humaines et de terres pour les besoins de la guerre et de l'impôt. Il y a plus de 4 ou 5000 ans, il existait déjà en Chine des descriptions chiffrées de la population et de l'agriculture.

Les Egyptiens de l'époque des Pharaons procédaient au dénombrement de la population ou du bétail.

A Rome, l'empereur Auguste fit procéder à une vaste enquête en dénombrant les soldats, les navires et les revenus publics.

Jusqu'au moyen âge, les seules « statistiques » existantes étaient les dénombrements faits dans des buts divers : assiettes de l'impôt, répartition des terres, recrutement dans l'armée sont effectuées avec des méthodes diverses (recensement des personnes, enregistrements de certains actes d'état civil….).

C'est à partir du XVIII siècle, qu'apparait le mot "statistique" crée par ACHENWAL en 1749 à partir du mot "STATISTA" (politique). Du simple dénombrement de populations humaines et de terres, la statistique est devenue

une science qui a retenu et continue de retenir l'attention, non seulement des empereurs et des rois, mais surtout des personnes de sciences.

L'extension et l'utilisation du calcul des probabilités développé par J.BERNOULLI au 18ème siècle et l'application des études démographiques et sociales ont permis à cette science de connaitre un essor considérable. Ainsi au 19ème siècle, de la simple statistique descriptive, elle passe au stade "Statistique Mathématique".

Depuis le 20ème siècle, les travaux de KARL PEARSON (1857-1936), de STUDENT (WILLIAM SEAL GOSSET, 1876-1937) et de RONAL FISCHER (1890-1963) ont permis à cette science de connaitre un développement considérable et une application vaste et variée. La statistique utilise les techniques et des méthodes de collecte, de présentation, d'étude et d'analyse des données quantitatives.

La statistique n'est pas uniquement utilisée pour décrire, pour mieux connaitre un évènement survenu dans le passé mais elle intervient de plus en plus dans les travaux de planification, dans le choix de prises de décisions et d'actions.

Bibliographie

1. A.CHEKROUN, statistiques descriptives et exercices, 2018
2. J.BLARD-LABORDERIE, l'essentiel des outils de statistique descriptive pour aborder des études en sciences humaines et sociales, 2015
3. G. CALOT, cours de statistique descriptive, Dunod, 1969.
4. G.CHAUVAT AND J.P REAU, statistique descriptive, Armand Colin, 2002.
5. M.TENENHAUS, statistique : Méthodes pour décrire, expliquer et prévoir, Dunod, 2006.
6. J,J DROESBEKE, éléments de statistiques, Ellipses, 2001.
7. L.LEBOUCHER and M-J VOISIN, introduction à la statistique descriptive, 2013
8. J.VAILLANT, éléments de statistique descriptive, 2015.

CHAPITRE I : INTRODUCTION A LA STATISTIQUE

I. Définition et concepts importants

1. Définition : le mot "**statistiques**" désigne un ensemble des données de faits numériques : recensement, statistiques des accidents, statistiques des emplois…. La "**statistique**" est une science qui consiste à recueillir, traiter et interpréter les données en vue d'en accroître les connaissances scientifiques, d'aider à la prise des décisions et de planifier les stratégies. A cet effet, elle utilise les méthodes mathématiques issues généralement du calcul des probabilités. Les données étudiées peuvent être de toute nature, ce qui rend la statistique utile dans tous les champs disciplinaires et explique pourquoi elle est enseignée dans toutes les filières universitaires, de la biologie à l'économie en passant par la géographie, la psychologie et bien sûr les sciences de l'ingénieur (agronomie…).

2. Concepts importants

La notion fondamentale en statistique est celle de groupe ou un ensemble d'objets équivalents que l'on appelle **population**. Ces objets sont appelés **individus**. L'étude de tous les objets d'une population finie s'appelle **recensement**. Lorsqu'on observe une partie seulement de la population, on parle de **sondage**. La partie étudiée est appelée **échantillon**. Chaque individu d'une population est décrit par un ensemble de caractéristiques appelées **variables** ou **caractères**. Ces variables présentent plusieurs modalités. Elles peuvent être classées selon leur nature en deux groupes:

- → Les ***variables qualitatives*** expriment l'appartenance à une catégorie ou une modalité. Elles ne se mesurent pas. Elles sont soit **nominales**, c'est-à-dire que leurs modalités ne peuvent pas être ordonnées, soit **ordinales** ; c'est-à-dire que leurs modalités possèdent une relation d'ordre.
- → Les ***variables quantitatives*** se mesurent et expriment une certaine quantité. Elles sont soit **discrètes**, c'est-à-dire que les valeurs prises par la variable sont en quantité finie ou dénombrables. Elles sont soit **continues**, c'est-à-dire qu'elles prennent des valeurs dans un intervalle des nombres réels.

  Exemple 1 :La couleur des cheveux est un caractère qui ne se mesure pas dont les modalités blonde, noire, blanche, ne possèdent pas d'ordre. C'est une variable qualitative nominale.

  Exemple 2 : dans la population des métaux, l'aluminium est un individu. Le caractère résistance prend les modalités « très résistant », « assez résistant », « peu résistant ». C'est une variable qualitative ordinale.

  Exemple 3 : le nombre des étudiants à l'ISMEA peut prendre des valeurs qui varient de 0 jusqu'à un nombre très important. C'est une variable quantitative discrète.

  Exemple 4 : la taille d'un être humain est une variable quantitative continue.

II. La démarche statistique

Elle consiste à traiter et interpréter les informations recueillies. Elle comporte deux grands aspects :

- → Aspect descriptif ou exploratoire
- → Aspect inférentiel ou décisionnel

1. La statistique exploratoire

Elle a pour but de synthétiser, résumer et structurer l'information contenue dans les données. Elle utilise pour cela des représentations des données sous forme des tableaux, des graphiques ou

d'indicateurs numériques. Connue sous le nom de statistique descriptive, cette phase est enrichie ses dernières années de nombreuses techniques de visualisation des données multidimensionnelles : c'est l'analyse des données. Le rôle de la statistique exploratoire est de mettre en évidence les propriétés de l'échantillon et de suggérer des hypothèses. Les principales méthodes d'analyse des données se séparent en deux groupes :

- → Les méthodes de classification visant à réduire la taille de l'ensemble en formant des groupes homogènes ;
- → Les méthodes factorielles qui cherchent à réduire le nombre des variables en résumant par un petit nombre.

2. La statistique inférentielle

Elle est l'ensemble des techniques visant à modéliser un ensemble de données en vue d'une extrapolation éventuelle à un ensemble plus vaste. Elle utilise de manière importante, les probabilités. Son but est d'étendre les propriétés constaté sur l'échantillon à la population toute entière et de valider ou d'infirmer les hypothèses à priori ou formulées après une phase exploratoire. Ces principaux éléments sont l'estimation, les tests, la corrélation et la régression :

- → L'estimation : elle permet à partir d'un ou de plusieurs échantillons de donner une valeur assez représentative d'une caractéristique de la variable (moyenne, écart-type…) qui sera considérer par la suite comme valable pour toute la population ;
- → Les tests : ayant observé un échantillon, on émet des hypothèses que l'on cherche ensuite par des méthodes appropriées à infirmer ou confirmer. Les tests statistiques permettent donc de prendre des décisions importantes. Citons par exemple le fait d'accepter ou non la livraison des pièces après avoir établi le taux des pièces défectueuses dans un échantillon.
- → La corrélation et la régression : elles permettent d'étudier la raison entre deux variables(ou plusieurs) et de prévoir leur comportement futur. Ce sont des outils de prévision.

3. Méthode statistique de résolution des problèmes

Pour résoudre un problème de statistique, les étapes suivantes sont fondamentales.

1) Identification du problème ou de la situation

Le staticien doit clairement identifier et définir la situation qui se présente à lui. Par exemple, l'insuffisance de production dans une usine, l'effet d'un produit alimentaire sur la population, le taux d'échec dans un établissement….

2) Rassemblement des données disponibles

Pour un problème donné, le staticien doit recueillir les données précises appropriées aussi complètes que possibles et pertinentes. Ces données peuvent provenir des sources internes (par exemple le service de production) ou des sources externes (diverses publications spécialisées). Dans ce dernier cas, il est préférable de recueillir les données de source primaire, c'est-à-dire des organismes ou agences qui ont initialement recueilli les données et les ont publiés les premiers plutôt que les sources secondaires, c'est-à-dire des organismes ou agences qui publient des données déjà parues auparavant.

3) Recueil de nouvelles données

Si les données disponibles ne sont pas suffisantes ou exhaustives de la situation, il est nécessaire de procéder à une collecte des nouvelles données. Cette collecte peut s'opérer grâce à des interviews par un questionnaire sur le terrain.

4) Classification et synthétisation des données

Une fois les données recueillies, il faut les classer ou les grouper afin de les rendre utilisables. Il est possible de synthétiser l'information contenue dans celles-ci pour en faciliter l'usage. Les tableaux, les graphiques et les indicateurs numériques sont des outils pour synthétiser l'information.

5) Analyse et interprétation des données

Une fois classer et synthétiser les données, il faut maintenant les analyser et les interpréter en vue :

- → D'en accroitre les connaissances scientifiques ;
- → De planifier des stratégies
- → D'aider à la prise de décision

4. Statistique et Géographie

Situées à l'un des grands carrefours des sciences naturelles, consacrées à l'étude de ces complexes dont les éléments simples sont toujours la terre et l'homme, comme les substances vivantes dans leur grande complexité ne renfermant que du carbone, de l'hydrogène, de l'oxygène et de l'azote, l'économique et la sociologie, la géographie et la statistique ne sauraient mener à bien ne nombreuses parties de leur œuvre sans travailler en commun. Mais la statistique et la géographie sont peut-être les plus intimement liées ; car si toutes les spécialités de la géographie recourent à la documentation statistique, la plus grande partie de la géographie humaine et notamment la géographie économique, serait presque impossible sans la statistique.

En pratique les utilisations de la statistique en géographie sont aussi que variées :

→ La statistique documente la géographie : les enquêtes des statisticiens établissent des documents nécessaires aux géographes. Au point de vue géographique, les statistiques sont la base des travaux de la géographie économique (statistiques agricoles, industrielles, douanières et sociales) et de la démographie (recensements de la population, hygiène publiques…) ; la géographie physique elle-même utilise de maintes statistiques surtout pour la climatologie, la minéralogie et l'océanographie ;

- → Représentation cartographique des travaux statistiques ;
- → Les cellules géographiques et les unités territoriales de la statistique ;
- → La classification et l'explication géographique.

En bref, la statistique est très largement utilisée en géographie surtout depuis les années 60 avec le développement de l'informatique. Les données que traite la géographie présentent certaines particularités : les individus statistiques sont fréquemment des unités spatiales, elles sont donc géo-localisables et cartographiables. Ces unités spatiales sont souvent des agrégats tels que l'ensemble d'habitants, d'entreprises etc. Ces outils statistiques permettent aux géographes de répondre à ses questions de prédilection comme: quels sont les principes de l'organisation de l'espace ? Existe-t-il des régularités, des spécificités locales ?

Exemple : étude d'une région dans un pays

5. Statistique et économie

La statistique représente un outil indispensable pour les responsables de la politique économique, le secteur économique privé. En effet, toutes les décisions concernant le monde économique à l'échelle nationale comme à l'échelle locale s'appuient sur les connaissances de l'évolution d'un certain de facteurs tels que : la structure des unités de production, la main d'œuvre, la démographie, les productions (en quantité et en qualité), les besoins, les résultats économiques des entités de production, les échanges économiques….

En pratique, les utilisations de la statistique dans le domaine économique sont aussi nombreuses que variées. Par exemple, la connaissance du revenu agricole par catégorie d'exploitation peut amener les responsables nationaux à prendre des mesures compensatrices telles que les subventions, les crédits agricoles…….

6. Statistique et agriculture

La statistique représente un outil indispensable pour les responsables des associations agricoles mais aussi les techniciens agricoles sur le terrain. En effet, toutes les décisions concernant le monde rural à l'échelle nationale comme à l'échelle locale s'appuient sur les connaissances de l'évolution d'un certain de facteurs tels que : la structure des unités de production, la main d'œuvre, la démographie, les productions (en quantité et en qualité), les besoins, les résultats économiques des exploitations agricoles, les échanges des produits agricoles et des produits nécessaires à la production agricole….

En pratique les utilisations de la statistique en agriculture sont aussi nombreuses que variées :

Exemples concernant la production agricole

- Superficie plantée en arachide au cours d'une campagne agricole. Cette estimation permet de mettre en œuvre les moyens nécessaires à la commercialisation, aux traitements, aux exportations….
- Dégâts causés par certains prédateurs. Le développement de certaines attaques inhabituelles peut par exemple amener les responsables agricoles à réorienter la production.

Exemples concernant la production animale

- Importance des cheptels : cette estimation permet de prévoir les doses de vaccins nécessaires pour les campagnes annuels de prophylaxie.
- Croissance du bétail : elle permet d'étudier l'adéquation entre l'offre et la demande.

Exemples concernant la foresterie

- Superficie des terres couvertes par la foret : leur connaissance précise permet de suivre l'évolution de la foret, éventuellement de quantifier la désertification et d'évaluer l'impact des campagnes de renouvellement ;
- L'importance des feux de brousse permet de prévoir les moyens à mettre en œuvre dans les différentes régions.

Exemples socio-économiques

- La connaissance du revenu agricole par catégorie d'exploitation
- L'évolution de ce revenu agricole peut amener les responsables nationaux à prendre des mesures compensatrices (moratoire, subventions, aides divers).
- Connaissance des habitudes alimentaires du consommateur : elle peut amener les responsables concernés à intensifier certaines cultures (maraîchères par exemple)

Outre son rôle fondamental dans la planification et donc la gestion, la statistique en agriculture tient un rôle important en ce qui concerne la recherche et l'expérimentation. Ainsi :

→ La comparaison des rendements d'une nouvelle variété et de variété préexistante ;
→ L'étude de l'efficacité de nouveaux pesticides ;
→ L'analyse et la comparaison de nouvelles techniques culturales.

7. Erreurs statistiques

→ ***L'obstacle des biais*** **:** un questionnaire mal formulé peut amener à des conclusions non valables. Les données sont alors biaisées. Il faut signaler que le facteur « frime » consistant à mettre l'accent sur les idées préconçues permet de transformer facilement les résultats réels en résultats désirés.

→ ***Les moyennes trompeuses :*** élucidons ce cas par un exemple. Sur 100 fermiers d'une région des USA, 99 ont un revenu moyen de 3 000 $ tandis que le dernier a un revenu de 1 000 000 $. L'un

des 99 fermiers veut vendre sa ferme et l'annonce dans un journal avec le commentaire suivant : dans la région le revenu moyen d'un fermier des de 10 000 $. A-t-il raison ? est-il honnête ?

→ ***La dispersion, la grande oubliée :*** l'exemple suivant illustre bien la situation. Un guerrier chinois menace ses troupes à la bataille rencontrant sur son chemin une rivière. Comme il n'avait pas de bateau et il savait que la profondeur de la rivière à cette période de l'année n'était que de 1 m en moyenne, le chinois ordonna à ses hommes de la traverser à pieds. Arrivés sur l'autre rive il constata avec stupeur que certains de ses hommes se sont noyés. Que s'est-il passé ?

→ ***Les embûches du*** "***post hoc ergo propter hoc***" c'est-à-dire "**après ceci, donc à cause de ceci**". C'est une expression latine pour exprimer le faut raisonnement qui dit "**puis que B suit A, alors B est causé par A".**

Exemple 1 : l'augmentation des arrivées des cargaisons de bananes dans le port de Douala a été suivie d'une augmentation du nombre de naissance à l'échelle nationale. Donc les bananes sont la cause de l'augmentation des naissances.

Exemple 2 : l'espérance de vie de l'homme a doublé dans le monde depuis la découverte du plant de tabac ; donc le tabac est source de longévité……

8. Rôle de l'ordinateur en statistique

L'ordinateur joue un rôle très important en statistique. Il peut être efficacement utilisé dans toute opération de traitement qui possède au moins une des caractéristiques suivantes :

→ Une grande quantité des données
→ Une répétition des projets
→ La nécessité d'une grande vitesse de traitement
→ La nécessité d'une grande précision
→ La complexité des opérations.

CHAPITRE II : DISTRIBUTION DES FREQUENCES ET REPRESENTATIONS GRAPHIQUES

I. Rangement des données et distributions des fréquences

I.1 Rangement des données

Supposons que les données brutes étaient recueillies selon le principe du chapitre introductif. Nous allons à présent les ranger suivant l'ordre croissant. Considérons l'exemple suivant : on a mesuré la taille des élèves du LCCL. On a obtenu le résultat suivant sur 50 élèves choisis au hasard : 1,22 ; 1,64 ; 1,03 ; 1,66 ; 1,29 ; 1,12 ; 1,49 ; 1,13 ; 1,33 ; 1,26 ; 1,57 ; 0,95 ; 1,62 ; 1,08 ; 1,72 ; 1,44 ; 1,28 ; 1,25 ; 1,59 ; 1,00 ; 1,56 ; 1,66 ; 1,48 ; 1,31 ; 1,35 ; 1,78 ; 1,53 ; 1,41 ; 1,66 ; 1,50 ; 1,30 ; 1,10 ; 1,20 ; 1,52 ; 1,36 ; 1,38 ; 1,33 ; 1,22 ; 1,23 ; 1,32 ; 1,23 ; 1,05 ; 1,23 ; 1,25 ; 1,70 ; 1,36 ; 1,42 ; 1,16 ; 0,94 ; 1,20.

Rangeons ces données par ordre croissant. A cet effet, construisons le diagramme "Steam and leaft".

0≤t<5		**5≥t**
4	0,9	5
03	1,0	85
032	1,1	6
0333202	1,2	96855
23013	1,3	5686
214	1,4	98
203	1,5	796
24	1,6	666
02	1,7	8

→

4	0,9	5
03	1,0	58
023	1,1	6
0022333	1,2	55689
0123	1,3	5668
124	1,4	89
023	1,5	679
24	1,6	666
02	1,7	8

On aura : **0,94 ; 0,95 ; 1,00 ; 1,03 ; 1,05 ; 1,08 ; 1,10 ; 1,12 ; 1,13 ; 1,16 ; 1,20 ; 1,20 ; 1,22 ; 1,22 ; 1,23 ; 1,23 ; 1,23 ; 1,25 ; 1,25 ; 1,26 ; 1,28 ; 1,29 ; 1,30 ; 1,31 ; 1,32 ; 1,33 ; 1,35 ; 1,36 ; 1,36 ; 1,38 ; 1,41 ; 1,42 ; 1,44 ; 1,48 ; 1,49 ; 1,50 ; 1,52 ; 1,53 ; 1,56 ; 1,57 ; 1,59 ; 1,62 ; 1,64 ; 1,66 ; 1,66 ; 1,66 ; 1,70 ; 1,72 ; 1,78.**

II.2 Distribution des fréquences

Lorsqu'on veut analyser une grande partie des données brutes, il est commode de les distribuer en classe ou catégorie et de déterminer le nombre d'individus appartenant à cette classe. Ce nombre est appelé **effectif** ou **amplitude** de la classe. Chaque effectif divisé par le nombre total d'individu est appelé **fréquence** de la classe. On obtient une distribution des effectifs ou des fréquences sous forme du tableau suivant :

Classes	Effectifs	Fréquences
[0,9 ; 1,0]	2	0,04
[1,0 ; 1,1]	4	0,08
[1,1 ; 1,2]	4	0,08
[1,2 ; 1,3]	12	0,24
[1,3 ; 1,4]	9	0,18
[1,4 ; 1,5]	5	0,10
[1,5 ; 1,6]	6	0,12
[1,6 ; 1,7]	5	0,10
[1,7 ; 1,8]	3	0,03
	50	**0,999~1**

Ainsi pour construire une distribution des fréquences, il est nécessaire de déterminer :

- Le nombre des classes à utiliser pour grouper les données ;
- La largeur de ces classes ;
- Les fréquences respectives des classes.

I.3 Quelques considérations pratiques

Lors de la construction d'une distribution des fréquences, il est souhaitable de respecter les règles suivantes :

a) Le nombre des classes utilisées pour occuper les données se situe généralement entre un minimum de 5 et un maximum de 15. Un très grand de classes est équivalent à un simple rangement des données tandis qu'un petit nombre ferait perdre beaucoup d'informations.

b) Les classes doivent être choisies de telle sorte que la plus petite et la plus grande observation appelées respectivement borne inferieure et borne supérieure y soient incluses et que chaque observation se trouve dans une et une seule classe.

c) De préférence, les largeurs des classes doivent être égales et s'il y'a lieu choisir des multiples de 5, 10, 50, 100…..

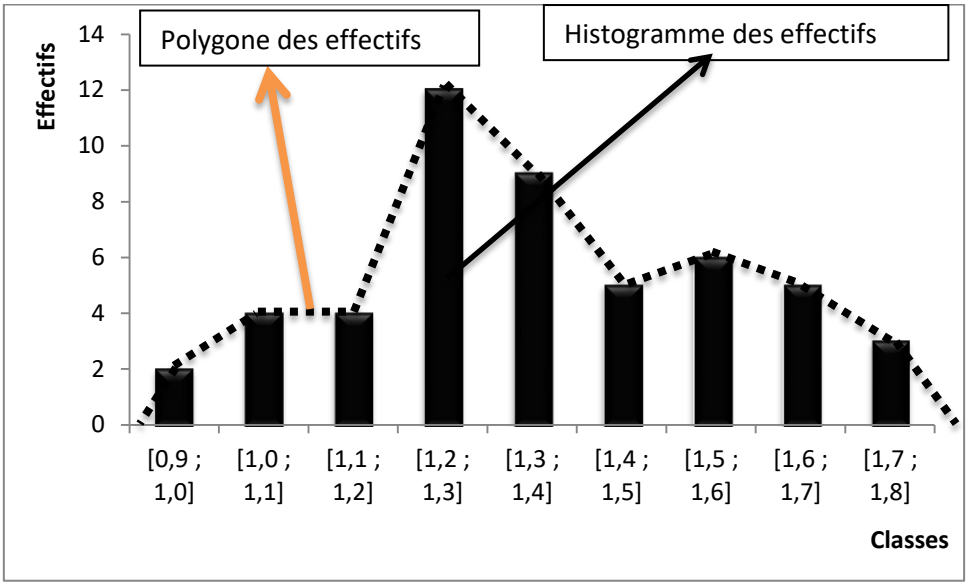
I.4 Quelques définitions

- On appelle ***étendue des données***, la différence entre la plus grande valeur et la plus petite valeur de l'observation des données. (1,78-0,94=0,84).
- On appelle ***étendue de la classe*** ou ***largeur de la classe***, la différence entre les bornes inférieure et supérieure de la classe. (1,0-0,9= 0,1)
- On appelle **centre de classe** la demi-somme des deux bornes de la classe ; c'est-à-dire la somme de la borne supérieure et de la borne inferieure divisée par 2.

II. Représentations graphiques

II.1 Histogramme

Un histogramme est une représentation graphique en tuyau de la distribution des fréquences ou des effectifs. Chaque classe est représentée par un rectangle dont la longueur correspond à l'amplitude et la largeur à l'étendue de la classe. Pour l'exemple précèdent, nous obtenons l'histogramme suivant :



II.2 Polygone des fréquences (ou des effectifs)

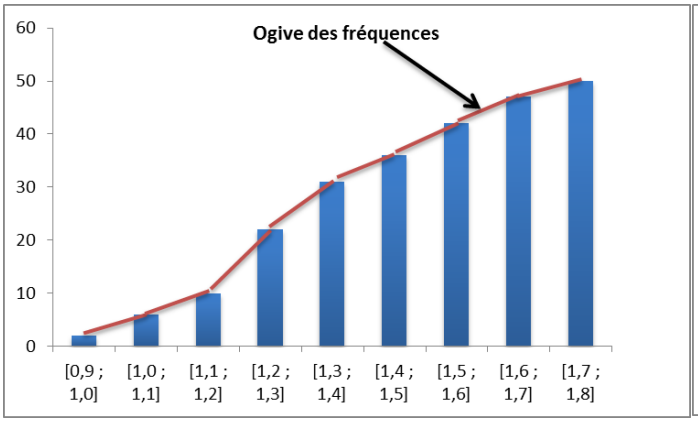
C'est une courbe qui s'obtient en joignant le centre des classes consécutives de l'histogramme des fréquences ou des effectifs. A l'origine puis à l'extrémité on crée une classe supplémentaire des mêmes étendues et d'amplitude nulle.

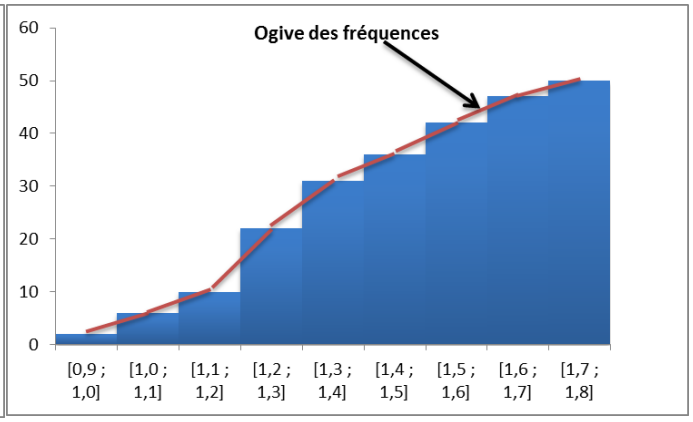
II.3 Ogive

C'est le polygone des fréquences cumulées ou des effectifs cumulés. Pour l'obtenir, on dresse d'abord la distribution des fréquences cumulées ou des effectifs cumulés. On construit en suite l'histogramme des fréquences cumulées ou des effectifs cumulés. L'ogive s'obtient alors en joignant le centre des classes de l'histogramme ainsi obtenu.

Distribution des effectifs cumulés et des fréquences cumulées.

Classes	**Effectifs**	**Fréquences**	**Effectifs cumulés**	**Fréquences cumulées**
[0,9 ; 1,0]	2	0,04	2	0,04
[1,0 ; 1,1]	4	0,08	6	0,12
[1,1 ; 1,2]	4	0,08	10	0,20
[1,2 ; 1,3]	12	0,24	22	0,44
[1,3 ; 1,4]	9	0,18	31	0,62
[1,4 ; 1,5]	5	0,10	36	0,72
[1,5 ; 1,6]	6	0,12	42	0,84
[1,6 ; 1,7]	5	0,10	47	0,94
[1,7 ; 1,8]	3	0,03	50	1
	50	**0,999~1**		





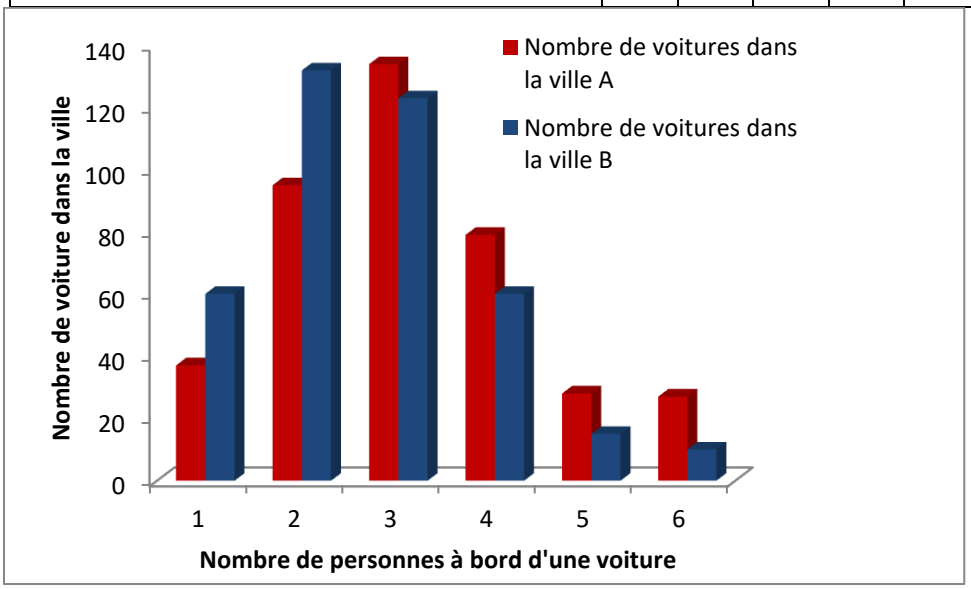
II.4 Autres diagrammes utilisés

A. Diagramme en bâtons (variable quantitative)

On appelle diagramme en bâtons (ou histogramme en bâtons) un graphique qui associe à chaque valeur de la variable un segment (bâton) dont la hauteur est proportionnelle à l'effectif.

Exemple : si on s'intéresse au nombre de personnes à bord d'une voiture dans deux villes différentes A et B, on peut dresser le tableau suivant :

Nombre de personnes à bord d'une voiture	1	2	3	4	5	6	Total
Nombre de voitures dans la ville A	37	95	134	79	28	27	400
Nombre de voitures dans la ville B	60	132	123	60	15	10	400



B. Diagramme en Camembert (variable qualitative)

On appelle diagramme en Camembert, un graphique qui divise un disque en secteur angulaire dont les angles au centre sont proportionnels aux effectifs de chaque modalité.

Pour une modalité donnée M_i d'effectif n_i, l'angle au centre α_i correspondant est donné (en degré Celsius) par la formule : $\alpha_i = n_i/n \times 360 = f_i \times 360$

Exemple :le tableau suivant donne les couleurs des cheveux d'un groupe de personnes.

Couleur des cheveux	Blond	Brun	Noir	Blanc	Total
Nombre de personnes présentant cette couleur	2365	2487	954	98	**5904**
Pourcentage des sujets (fréquences)	40	42,1	16,2	1,7	**100**

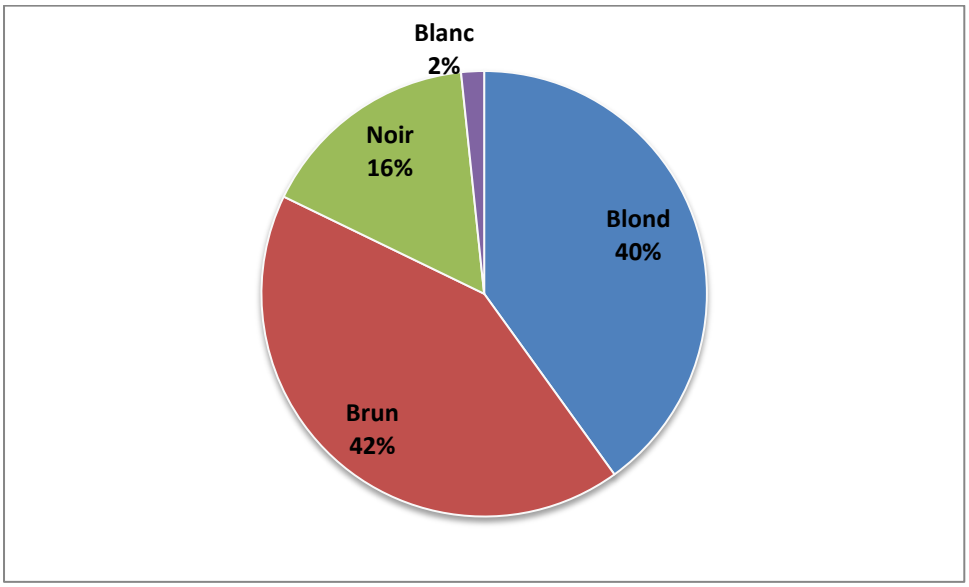
Calcul des angles au centre de chaque secteur

Couleur des cheveux	Blond	Brun	Noir	Blanc
Angles des secteurs en degré	144°	151,1°	58,3°	6,1°

Avec :

$\alpha_1 = 40/100 \times 360$ $\alpha_2 = 42,1/100 \times 360$ $\alpha_3 = 16,2/100 \times 360$

$\alpha_4 = 1,7/100 \times 360$



CHAPITRE III: CARACTERISTIQUES DES DONNES UNIDIMENSIONNELLES

I. Caractéristiques de tendance centrale

Dans plusieurs cas les données ont tendance à se rassembler autour d'une valeur centrale et celle-ci est souvent utilisée pour décrire l'aspect général des données. L'objet de ces mesures est de résumer en une seule valeur la grandeur typique, le milieu ou le centre des données. La plus familière de ces mesures est la moyenne arithmétique. La médiane et le mode constituent d'autres mesures de tendance centrale.

1.1 Moyenne arithmétique

1.1.1 Définition

On appelle ***moyenne arithmétique*** de n données numériques $x1, x2, \dots\ xn$, le nombre noté $\overline{X}$ défini par : $\overline{X} = x1 + x2 + \cdots \frac{xn}{n} = \frac{1}{n}\sum xi$

Souvent, on associe aux nombres $x1, x2, \dots\ xn$ des facteurs d'importance ou poids $w1, w2, \dots\ wn$dépendants de la signification ou de l'importance que l'on donne à ces nombres.

$$\overline{X} = w1x1 + w2x2 + \cdots wn\ xn/w1 + w2 + \cdots wn = \sum wn/\sum w$$

Exemple : dans le calcul de la moyenne du passage au niveau supérieur à l'ISMEA, on utilise la moyenne pondérée avec les coefficients représentant les matières.

1.1.2 Méthode pratique de calcul de la moyenne arithmétique

La moyenne arithmétique pour les données groupées s'obtient par la formule suivante :

$$X = \frac{\sum_{i=1}^{k} fi.mi}{n}$$

$fi: fréquences\ au\ nombre\ d'observation\ dans\ la\ classe\ i$

$mi: centre\ de\ classe\ i$

$k: nombre\ total\ de\ classe$

$n: taille\ des\ données$

Exemple : reprenons l'exemple du chapitre précèdent.

Classes	mi	Effectifs	Fréquences
[0,9 ; 1,0]	0,95	2	0,04
[1,0 ; 1,1]	1,05	4	0,08
[1,1 ; 1,2]	1,15	4	0,08
[1,2 ; 1,3]	1,25	12	0,24
[1,3 ; 1,4]	1,35	9	0,18
[1,4 ; 1,5]	1,45	5	0,10
[1,5 ; 1,6]	1,55	6	0,12
[1,6 ; 1,7]	1,65	5	0,10
[1,7 ; 1,8]	1,75	3	0,03

$$X = \frac{2x0,95 + 4x1,05 + 4x1,15 + 12x1,25 + 9x1,35 + 5x1,45 + 6x1,55 + 5x1,65 + 3x1,75}{50}$$

$$\boldsymbol{X = 1,358}$$

Idem avec les fréquences x=1,358

1.2 Médiane

1.2.1 Définition

La médiane représente la valeur qui occupe la place du milieu dans le rangement ascendant ou descendant des données. C'est la valeur **Me/ F (Me)= 0,5**où F est la fonction de répartition ou fonction des fréquences cumulées.

1.2.2 Calcul pratique de la médiane

- Pour les données non groupées, si les observations sont rangées par ordre croissant alors :

$$\begin{cases} Me = \frac{X_{n+1}}{2} \text{ Si n est impair} \\ Me = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2} \text{ Si n est pair} \end{cases}$$

- Lorsque les données sont représentées en classes, on détermine la classe médiane $[e_{i-1} ; e_i[$ tel que $\mathbf{F_{i-1} = F(e_{i-1}) \leq 0,5}$ et $\mathbf{Fi = F(e_i) > 0,5}$. **F = fonction des fréquences cumulées.**

La médiane est alors obtenue par la formule suivante :

$$Me = \mathbf{e_{i-1} + a_i \times \frac{0,5 - Fi - 1}{fi}}$$

$\mathbf{e_{i-1}}$ **:** borne inférieure de la classe médiane ;

$\mathbf{a_i}$ **:** largeur de la classe médiane

$\mathbf{F_{i-1}}$ **:** fréquences cumulées de toutes les classes ***précédant*** la classe médiane ;

$\mathbf{f_i}$ **:** fréquence de la classe médiane.

Exemple : Reprenons l'exemple du chapitre précédant.

- Calculons la médiane pour les données groupées. N= 50 donc n pair

$$Me = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2} = \frac{X_{\frac{50}{2}} + X_{\frac{50}{2}+1}}{2} = \frac{X25 + X26}{2} = \frac{1,32 + 1,33}{2} = 1,325$$

Me = 1,325

- Calculons la médiane pour les données groupées

Me = $\mathbf{e_{i-1} + a_i \times \frac{0,5 - Fi - 1}{fi}}$ avec [1,3 ; 1,4[la classe médiane. $\mathbf{e_{i-1} = 1,3 ; a_i = 0,1 ; f_i = 0,18 ; F_{i-1} = 0,44}$

⟶ Me= 1,03+0,1 x $\frac{0,5-0,44}{0,18}$= 1,333 **Me = 1,333**

Remarque : Si aucune précision n'est donnée, on retiendra la médiane issue des données groupées.

1.3 Mode

1.3.1 Définition et exemple

Le mode d'un ensemble des nombres est le nombre qu'on rencontre le plus fréquemment c'est-à-dire celui qui a la plus grande fréquence. Le mode peut ne pas exister et s'il existe, il ne peut pas être unique.

Exemples :

- La série (1 ; 2 ; 4 ; 3 ; 3 ; 7 ; 8 ; 8 ; 8 ; 9 ; 11 ; 11) a pour mode 8. Mo = 8
- La série (1 ; 2 ; 2 ; 2 ; 3 ; 3 ; 7 ; 8 ; 8 ; 8 ; 11 ; 11) a deux modes : 2 et 8. M_{o1}= 2 et M_{o2}= 8
- La série (1 ; 2 ; 4 ; 3 ; 7 ; 8 ; 9 ; 11) n'a pas de mode.

1.3.2 Calcul pratique du mode

A partir de la distribution des fréquences, la formule qui suit nous indique comment évaluer le mode.

$$Mo = e_{j-1} + a_j \times \frac{Dj-1}{Dj-1+Dj}$$

Où :

e_{j-1} : borne inférieure de la classe modale

a_j : largeur de la classe modale

D_{j-1} : différence entre la fréquence de la classe modale et de la classe ***précédente*** dans la distribution ;

D_j: différence entre la fréquence de la classe modale et de la classe ***suivante*** dans la distribution ;

Calcul pratique

Avec [1,2 ; 1,3[est la classe modale. **e_{j-1} = 1,2 ; a_j= 0,1 ; D_{j-1} = 12-4 ; D_j= 12-9**

$$Mo = 1,2+0,1 \times \frac{12-4}{(12-4)+(12-9)} = 1,2727 \quad \mathbf{Mo = 1,273}$$

1.4 Quartiles-Déciles-Quantiles

Si un ensemble de nombres est rangé par ordre de grandeur croissante, le nombre divisant l'ensemble en 2 parties égales est ***la médiane***. Par extension, on peut penser aux valeurs qui divisent l'ensemble en 4 parties égales. On note ces **Q_1, Q_2,** et **Q_3** et on les appelle respectivement le ***1er, 2ème et 3ème quartile***. **Q_2**étant la médiane.

De même, on appelle ***déciles***, les valeurs qui divisent les données en 10 parties égales et on les note **: D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8** et **D_9** tandis que les valeurs qui divisent les données en 100 parties égales sont appelées ***quantiles d'ordre 100*** ou ***centiles*** et notées **P_1, P_2, …..P_{99}**.

Le 5ème décile et le 50ème centile correspondent à la médiane. Le 25ème et 75ème centile correspondent respectivement aux **Q_1**et**Q_3.**

Exemple : Déterminons les quartiles et les déciles pour la distribution suivante :

- Calculons **Q_1**

  Q_1est / $F_{i-1} \leq 0,25$ et $f_i > 0,25 \rightarrow Q_1 = e_{i-1} + a_i \times \frac{0,25-Fi-1}{fi}$

  La classe de Q_1est [1,2 ; 1,3[$\rightarrow Q_1 = 1,2 + 0,1x \frac{0,25-0,20}{0,24} \rightarrow \underline{Q_1 = 1,220}$

- Calculons **$Q_2 \rightarrow Q_2$= Me**

- Calculons Q_3

Q_3 est / $F_{i-1} \leq 0{,}75$ et $f_i > 0{,}75 \rightarrow Q_{1=} e_{i-1} + a_i \times \frac{0{,}75 - Fi - 1}{fi}$

La classe de Q_3est [1,5 ; 1,6[$\rightarrow Q_{1=} 1{,}5 + 0{,}1 \times \frac{0{,}75 - 0{,}72}{0{,}12} \rightarrow \underline{Q_{3=} 1{,}525}$

Calcul des déciles

Déterminons les déciles

- D_1 est / $F_{i-1} \leq 0{,}1$ et $f_i > 0{,}1 \rightarrow D_{1=} e_{i-1} + a_i \times \frac{0{,}1 - Fi - 1}{fi}$

La classe de D_1est [1,0 ; 1,1[$\rightarrow D_{1=} 1{,}0 + 0{,}1 \times \frac{0{,}10 - 0{,}04}{0{,}08} \rightarrow \underline{D_{1=} 1{,}075}$

- D_2 est / $F_{i-1} \leq 0{,}2$ et $f_i > 0{,}2 \rightarrow D_{2=} e_{i-1} + a_i \times \frac{0{,}2 - Fi - 1}{fi}$

La classe de D_2est [1,2 ; 1,3[$\rightarrow D_{2=} 1{,}2 + 0{,}1 \times \frac{0{,}2 - 0{,}20}{0{,}24} \rightarrow \underline{D_{2=} 1{,}2}$

- D_3 est / $F_{i-1} \leq 0{,}3$ et $f_i > 0{,}3 \rightarrow D_{2=} e_{i-1} + a_i \times \frac{0{,}3 - Fi - 1}{fi}$

La classe de D_3est [1,2 ; 1,3[$\rightarrow D_{3=} 1{,}2 + 0{,}1 \times \frac{0{,}3 - 0{,}20}{0{,}24} \rightarrow \underline{D_{3=} 1{,}241}$

- D_4 est / $F_{i-1} \leq 0{,}4$ et $f_i > 0{,}4 \rightarrow D_{4=} e_{i-1} + a_i \times \frac{0{,}4 - Fi - 1}{fi}$

La classe de D_2est [1,2 ; 1,3[$\rightarrow D_{4=} 1{,}2 + 0{,}1 \times \frac{0{,}4 - 0{,}2}{0{,}24} \rightarrow \underline{D_{4=} 1{,}283}$

- D_5 est / $F_{i-1} \leq 0{,}5$ et $f_i > 0{,}5 \rightarrow D_{2=} e_{i-1} + a_i \times \frac{0{,}5 - Fi - 1}{fi} \rightarrow D_5 = Me = 1{,}333$

- D_6 est / $F_{i-1} \leq 0{,}6$ et $f_i > 0{,}6 \rightarrow D_{4=} e_{i-1} + a_i \times \frac{0{,}6 - Fi - 1}{fi}$

La classe de D_6est [1,3 ; 1,4[$\rightarrow D_{6=} 1{,}3 + 0{,}1 \times \frac{0{,}6 - 0{,}4}{0{,}18} \rightarrow \underline{D_{6=} 1{,}389}$

- D_7 est / $F_{i-1} \leq 0{,}7$ et $f_i > 0{,}7 \rightarrow D_{7=} e_{i-1} + a_i \times \frac{0{,}7 - Fi - 1}{fi}$

La classe de D_7est [1,4 ; 1,5[$\rightarrow D_{7=} 1{,}4 + 0{,}1 \times \frac{0{,}7 - 0{,}62}{0{,}10} \rightarrow \underline{D_{7=} 1{,}48}$

- D_8 est / $F_{i-1} \leq 0{,}8$ et $f_i > 0{,}8 \rightarrow D_{8=} e_{i-1} + a_i \times \frac{0{,}8 - Fi - 1}{fi}$

La classe de D_8est [1,5 ; 1,6[$\rightarrow D_{8=} 1{,}5 + 0{,}1 \times \frac{0{,}8 - 0{,}72}{0{,}12} \rightarrow \underline{D_{8=} 1{,}566}$

- D_9 est / $F_{i-1} \leq 0{,}9$ et $f_i > 0{,}9 \rightarrow D_{9=} e_{i-1} + a_i \times \frac{0{,}9 - Fi - 1}{fi}$

La classe de D_9est [1,6 ; 1,7[$\rightarrow D_{9=} 1{,}6 + 0{,}1 \times \frac{0{,}9 - 0{,}84}{0{,}10} \rightarrow \underline{D_{9=} 1{,}66}$

II. Caractéristiques de dispersion

2.1 L'étendue

Soit$x1, x2, \ldots xn$, une série statistique que l'on suppose ordonnée. L'étendue est la différence entre les valeurs extrêmes : $\mathbf{W = Xn - X1}$ **=1,78-0,94.**

2.2 L'interquartile

Les quartiles Q_1, Q_2, et Q_3étant définis, la différence Q_3-Q_1 appelée ***interquartile*** est un indicateur pour mesurer la dispersion. Parfois, on utilise la quantité ½ (Q_3 - Q_1) appelée ***intervalle semi-interquartile***.

2.3. Le diagramme en boîte

Le diagramme en boîte ou **"box plot"** représente schématiquement les principales caractéristiques d'une variable numérique en utilisant les quartiles. Dans sa version la plus courante, la partie centrale est représentée par une boite de largeur arbitraire et dont la longueur correspond à l'interquartile. On trace à l'intérieur la position de la médiane. La boîte est alors complétée par "***les moustaches***" correspondant aux valeurs adjacentes.

- Valeur adjacente inferieure :c'est la plus petite valeur **> Q_1- 1,5 (Q_3 - Q_1) ;**
- Valeur adjacente supérieure : c'est la plus grande valeur **< Q_3 +1,5 (Q_3 - Q_1).**

Considérons l'exemple précèdent. On a **Q_1= 1,220 ; Q_2= 1,333 ; Q_3= 1,525**

AI= Q_1- 1,5 (Q_3 - Q_1) = 1,220 – 1,5 (1,525 -1,220) = 0,762

AS = Q_3 +1,5 (Q_3 - Q_1) = 1,525 + 1,5 (1,525 -1,220) = 1,982

2.4 L'écart-type et la variance

2.4.1 Définition

L'écart-type est la mesure de dispersion la plus utilisée. Pour n données$x1, x2, \dots\ xn$, l'écart-type s'obtient par la formule suivante : $\mathcal{S} = \frac{\sqrt{\sum_{i=1}^{n}(xi-\dot{X})^2}}{n}$

Remarque : l'écart-type est aussi appelé **moyenne quadratique.**

2.4.2 Calcul pratique de l'écart-type

Pour les données groupées, on utilise la formule suivante pour calculer l'écart-type :

$$\mathcal{S} = \frac{\sqrt{\sum_{i=1}^{k} fi(mi-\dot{X})^2}}{n}$$

Avec : $n : taille\ des\ données$; $k : nombre\ de\ classe$; $fi : fréquence\ ou\ effecetif\ de\ la\ classe\ i$; $mi : centre\ de\ la\ classe\ i$

Exemple : Calculons l'écart-type des données du chapitre précèdent

Classes	mi	**Effectifs**	$mi - X$	$(mi - X)^2$	$fi(mi - X)^2$
[0,9 ; 1,0]	0,95	2	-0,408	0,166	0,3329
[1,0 ; 1,1]	1,05	4	-0,308	0,094	0,3794
[1,1 ; 1,2]	1,15	4	-0,208	0,043	0,1730
[1,2 ; 1,3]	1,25	12	-0,108	0,011	0,1399
[1,3 ; 1,4]	1,35	9	-0,008	0,0000	0,0000
[1,4 ; 1,5]	1,45	5	0,092	0,008	0,042
[1,5 ; 1,6]	1,55	6	0,192	0,036	0,2211
[1,6 ; 1,7]	1,65	5	0,292	0,085	0,4263
[1,7 ; 1,8]	1,75	3	0,392	0,153	0,4609
					∑= 2,1755

$$\mathcal{S} = \sqrt{2,1755}/50 = 0,208$$

2.4.3 La variance

La variance est aussi une mesure de dispersion des données. Elle est égale au carré de l'écart-type et est notée $\mathcal{s}2$.

Exemple : $\mathcal{s} = 0{,}208 \rightarrow \mathcal{s}2 = 0{,}0432$

2.5 Dispersion absolue et relative. Coefficient de variation

L'écart-type ou toute autre mesure de dispersion s'exprimant dans l'unité de mesure des données est appelée dispersion absolue. On définit la dispersion relative comme suit : ***dispersion relative= dispersion absolue/moyenne.***

Si la dispersion absolue est l'écart-type $\mathcal{s}$ et la moyenne est la moyenne arithmétique X alors la dispersion relative est appelée coefficient de variation et noté CV avec **CV=** $\mathcal{s}/\boldsymbol{X}$. Le CV est indépendant du choix des unités des données.

La dispersion absolue est fonction des unités de mesure des données or la dispersion relative est indépendante des mesures des données.

CHAPITRE IV: DESCRIPTION BIDIMENSIONNELLE DES DONNEES
CARACTERISTIQUES DES DONNES BIDIMENSIONNELLES

4.1 Introduction

Dans ce chapitre, nous introduisons quelques outils utiles à l'étude et à l'examen de relations et de liens entre différentes variables sur une même population. Le cas le plus simple est celui de deux variables ***X*** et ***Y***. souvent, la variable ***X*** est **"manipulable"** par l'expérimentateur : il peut s'agir par exemple du dosage d'un traitement, ou du sexe des personnes que l'on choisit d'interroger. On l'appelle en sciences humaines une ***variable indépendante***.

L'autre variable ***(Y)*** est alors appelée ***variable dépendante***. Par exemple si X est le dosage d'un traitement, Y peut être l'intensité de la douleur manifestée par un malade, et si X est le sexe des personnes interrogées, Y peut être tout à fait leur taille.

La variable Y est parfois aussi appelée ***variable prédite***, sous-entendant une relation de cause à effet entre X et Y. on cherche ainsi à prédire Y en fonction de X, c'est-à-dire à établir une relation du type ***Y= f (X)*** entre les deux variables. Les relations les plus simples à étudier sont les relations affines que nous appellerons aussi relations linéaires.

Pour ce faire nous aborderons deux notions dans ce chapitre :

→ **La corrélation** qui mesure l'intensité du lien entre deux variables
→ **La régression linéaire** qui est la relation de prédiction (affine) entre les deux variables.

Lorsque nous rassemblons des observations sur plusieurs variables statistiques sur une même population, nous avons souvent besoin, avant l'élaboration d'une analyse fine, de représentations graphiques qui nous donneront une première impression sur l'intensité du lien (qu'on doit confirmer par des méthodes plus élaborées) entre deux variables.

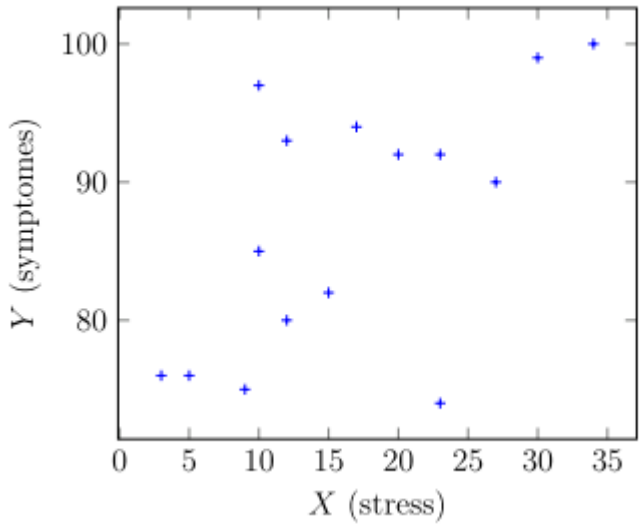
4.2 Nuage statistique

Pour un couple de variables (X ; Y) chaque individu est représenté dans le plan par un point dont les coordonnées (xi ;yi) sont les mesures de X et Y pour cet individu. L'ensemble de ces points s'appelle le nuage statistique ou le diagramme de dispersion. Les deux variables X et Y ont souvent des rôles distincts : lorsqu'une des deux variables (la variable dite indépendante) est manipulable par l'expérimentateur on la note X et on la représente horizontalement (en abscisse). L'autre variable (notée généralement Y) est représentée verticalement (en ordonnée).

4.2.1 Exemple

Des chercheurs ont étudié la relation entre le stress et la santé mentale. Ils ont mis au point une échelle qui permet de donner une mesure du stress pour chaque personne interrogée. Ils ont demandé également aux personnes interrogées de remplir la liste de contrôle d'Hopkins, qui évalue la présence ou l'absence d'un certain nombre de symptômes psychologiques. Le tableau suivant représente les mesures du stress X et des symptômes Y pour 15 personnes.

	A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*	*M*	*N*	*O*
X	***12***	***30***	***27***	***9***	***20***	***3***	***12***	***15***	***5***	***10***	***23***	***34***	***23***	***10***	***17***
Y	***80***	***99***	***90***	***75***	***92***	***76***	***93***	***82***	***76***	***85***	***74***	***100***	***92***	***97***	***94***



Sur ce nuage de point on remarque que la majorité des points sont assez proches d'une droite qui est la diagonale du carré délimitant la figure. On peut donc déjà s'attendre, au vu de la figure, à trouver un lien fort entre les variables et que ce lien entre les variables soit linéaire.

4.3 Coefficients de Corrélation

Dans cette partie nous aborderons deux coefficients de corrélations différents :

→ **Le coefficient de corrélation linéaire (ou "coefficient de corrélation de Pearson")** qui traduit le fait que deux variables soient liées par une relation linéaire (ou affine), c'est-à-dire le fait que les points du nuage statistique soient concentrés autour d'une droite ;

→ **Le coefficient de corrélation des rangs de Spearman** qui traduit le fait qu'une des variables augmente (ou diminue) quand l'autre augmente. Dans l'exemple précèdent, ce coefficient permettrait de confirmer que les symptômes augmentent quand le stress augmente. Le calcul du coefficient de corrélation linéaire s'appuiera sur un paramètre statistique appelé covariance.

4.3.1 Covariance

Par définition la covariance de deux variables X et Y est la moyenne des produits des écarts des deux variables. Ce qui donne la définition mathématique suivante :

$$\mathbf{Cov\ (X\ ;\ Y)= m\ (\ (X-m(X)\ )\ (Y-m(Y))\ )}$$

En pratique, on utilisera une définition plus simple à mettre en œuvre et rigoureusement équivalente :

$$\mathbf{Cov\ (X\ ;\ Y)= m\ (XY)\ -\ m(X)\ m(Y)}$$

Sur des petits échantillons de taille n la covariance s'obtient en faisant les calculs suivants :

$$m(X) = \frac{\sum x_i}{n}, \qquad m(Y) = \frac{\sum y_i}{n}, \qquad m(XY) = \frac{\sum x_i y_i}{n},$$

$$cov(X;Y) = \frac{\sum x_i y_i}{n} - \left(\frac{\sum x_i}{n}\right)\left(\frac{\sum y_i}{n}\right)$$

Calcul pour l'exemple précédent :

Calculs pour l'exemple 2.2.1 défini précédemment :

$$m(X) = \frac{\sum x_i}{n} = \frac{12+30+\cdots+17}{15} = \frac{250}{15} \simeq 16{,}67.$$

$$m(Y) = \frac{\sum y_i}{n} = \frac{80+99+\cdots+94}{15} = \frac{1305}{15} = 87.$$

$$m(XY) = \frac{12\times80+30\times99+27\times90+\cdots+17\times94}{15} = \frac{22465}{15} \simeq 1497{,}667.$$

$$Cov(X;Y) = m(XY) - m(X)\,m(Y) = \frac{22465}{15} - \frac{250}{15}\frac{1305}{15} \simeq 47{,}667.$$

4.3.2 Le coefficient de corrélation linéaire de Pearson

Par définition, le coefficient de corrélation est le rapport entre la covariance et le produit des écarts-types.

$$r(X;Y) = \frac{Cov(X;Y)}{s(X)s(Y)}.$$

Calculs pour l'exemple 2.2.1 défini précédemment :

$$m(X^2) = \frac{\sum x_i^2}{N} = \frac{12^2+30^2+\cdots+17^2}{15} = \frac{5360}{15}.$$

$$Var(X) = m(X^2) - m(X)^2 = \frac{5360}{15} - \left(\frac{250}{15}\right)^2 \simeq 79{,}56.$$

$$s(X) = \sqrt{Var(X)} \simeq 8{,}92.$$

$$m(Y^2) = \frac{\sum y_i^2}{N} = \frac{80^2+99^2+\cdots+94^2}{15} = \frac{114725}{15}.$$

$$Var(Y) = m(Y^2) - m(Y)^2 = \frac{114725}{15} - \left(\frac{1305}{15}\right)^2 \simeq 79{,}33.$$

$$s(Y) = \sqrt{Var(Y)} \simeq 8{,}91.$$

$$\text{D'où } r(X;Y) = \frac{Cov(X;Y)}{s(X)s(Y)} = \frac{47{,}667}{8{,}92\times8{,}91} \simeq 0{,}6.$$

<u>Interprétations et remarques</u>

1. Le coefficient de corrélation est compris entre -1 et +1 ;
2. On dira qu'on a une corrélation très forte (positive ou négative) si $|r(X;Y)| \geqslant 0{,}75$. Dans ce cas on doit s'attendre à ce que chaque variable soit un bon prédicteur pour l'autre.
   a) Si r est positif, le lien entre les variables X et Y signifie que Y augmente linéairement quand X augmente ;
   b) Si r est négatif, le lien entre les variables X et Y signifie que Y diminue linéairement quand X augmente.

TRAVAUX DIRIGES DE STATISTIQUE DESCRIPTIVE

Exercice 1

On veut étudier la valeur du dosage d'une certaine protéine dans le sang. Pour cela, on dispose de 40 dosages exprimés en g/l à savoir :

1,09 ; 0,94 ; 1,18 ; 1,12 ; 0,75 ; 0,97 ; 1,27 ; 1,33 ; 1,10 ; 1,04 ; 0,85 ; 0,88 ; 1,26 ; 1,06 ; 1,03 ; 0,80 ; 0,74 ; 1,03 ; 1,19 ; 1,25 ; 0,89 ; 1,35 ; 0,91 ; 0,91 ; 1,07 ; 1,14 ; 0,72 ; 1,03 ; 1,24 ; 0,98 ; 1,18 ; 1,23 ; 0,77 ; 1,09 ; 1,05 ; 1,18 ; 1,25 ; 0,83 ; 1,13 ; 0,83.

1. Faire un diagramme « Steam and Leaf » afin d'ordonner ces données par ascendance.
2. Combien y'a-t-il de possibilités de regrouper ces données en classes de largeur « identique » ? indiquer à chaque fois le nombre de classes et la largeur.
3. On choisit de regrouper ces données en classes en optant pour un nombre minimal de classes issu de la question 2. Donner le tableau de la distribution en classes en précisant le centre des classes, les effectifs, les fréquences, les effectifs cumulés et les fréquences cumulées.
4. Calculer la moyenne, la médiane et le mode : pour les données brutes et pour les données groupées.

Exercice 2

On a demandé aux étudiants de l'IUSAE de remplir un questionnaire d'évaluation des cours. Cinq (05) catégories de réponses sont possibles. L'une des questions posées était : comparé aux autres cours que avez suivi, quelle est la qualité générale du cours de statistique ? Mauvaise, Equivalente, Bonne, Très Bonne, Excellente.

Les réponses d'un échantillon de **soixante (60) étudiants** qui ont suivi le cours de statistique sont les suivantes. Pour faciliter le traitement informatique des résultats du questionnaire, un code numérique a été utilisé : **1=Mauvaise ; 2= Equivalente ; 3= Bonne ; 4= Très Bonne et 5= Excellente.**

3 4 4 5 1 5 3 4 5 2 4 5 3 4 4 4 5 5 4 1 4 5 4 2 5 4 2 4 4 4
5 5 3 4 5 5 2 4 3 4 5 4 3 5 4 4 3 5 4 5 4 3 5 3 4 4 3 5 3 3

a) Quelle est la variable étudiée ? Donner sa nature
b) Déterminer la moyenne, le mode et la médiane de cette série.
c) Résumer les données dans un tableau contenant les modalités, les effectifs et les fréquences et calculer le mode de ces données.

Exercice 3

Chez un fabricant de tubes de plastiques, on a prélevé un échantillon de 100 tubes dont on a mesuré le diamètre en décimètre.

1,94	**2,20**	**2,33**	**2,39**	**2,45**	**2,50**	**2,54**	**2,61**	**2,66**	**2,85**
1,96	**2,21**	**2,33**	**2,40**	**2,46**	**2,51**	**2,54**	**2,62**	**2,68**	**2,87**
2,07	**2,26**	**2,34**	**2,40**	**2,47**	**2,52**	**2,55**	**2,62**	**2,68**	**2,90**
2,09	**2,26**	**2,34**	**2,40**	**2,47**	**2,52**	**2,55**	**2,62**	**2,68**	**2,91**
2,09	**2,28**	**2,35**	**2,40**	**2,48**	**2,52**	**2,56**	**2,62**	**2,71**	**2,94**
2,12	**2,29**	**2,36**	**2,41**	**2,49**	**2,52**	**2,56**	**2,63**	**2,73**	**2,95**
2,13	**2,30**	**2,37**	**2,42**	**2,49**	**2,53**	**2,57**	**2,63**	**2,75**	**2,99**
2,14	**2,31**	**2,38**	**2,42**	**2,49**	**2,53**	**2,57**	**2,65**	**2,76**	**2,99**
2,19	**2,31**	**2,38**	**2,42**	**2,49**	**2,53**	**2,59**	**2,66**	**2,77**	**3,09**
2,19	**2,31**	**2,38**	**2,42**	**2,50**	**2,54**	**2,59**	**2,66**	**2,78**	**3,12**

1. Identifier la population, les individus, le caractère étudié et son type.
2. Déterminer le mode et la médiane des données.
3. Dresser le tableau de distribution des effectifs, des fréquences, des effectifs cumulés, des fréquences cumulées avec des classes d'amplitude 0,15.
4. Tracer l'histogramme de cette variable statistique.
5. Déterminer la valeur du diamètre au-dessous de laquelle se trouvent 50% des tubes de plastique.
6. Déterminer le pourcentage de tubes ayant un diamètre inférieur à 2,58.

Exercice 4

Le tableau ci-dessous représente les notes obtenues à un devoir par une classe de 32 élèves.

Notes	5	6	7	8	9	10	11	12	13	14	15
Effectifs	1	3	2	4	3	9	4	1	2	1	2

Le professeur principal décide d'analyser cette série statistique afin de mieux les analyser. Il vous demande donc de lui déterminer le nombre de notes inférieures à 10, le nombre de notes supérieures ou égales à 10, l'étendue, la médiane, le mode et la moyenne arithmétique de ces notes.

Déterminer le premier quartile $\mathbf{Q_1}$ et le troisième quartile $\mathbf{Q_3}$ de la série, puis interpréter ces résultats.

Exercice 5

A la question ***"Les statistiques permettent de mentir avec assurance : Quelle est votre opinion ?",*** 80 personnes interrogées ont ainsi répondu :

Pas du tout d'accord	10
Plutôt d'accord	15
Indifférente	12
Plutôt en accord	18
Tout à fait d'accord	25

Soit *X* la variable associée à cette enquête.

1. Quelles sont les modalités de X ? Quel est son type ? Quel est son mode ?
2. Etablir une distribution des fréquences de cette variable et représenter là par un diagramme circulaire.
3. Quelle est la proportion des personnes n'ayant pas d'opinion tranchée sur la question ? Ayant une opinion extrêmement tranchée sur la question ?

Exercice 6:

Une pépinière prépare Noël et mesure les 400 sapins qu'elle a déterrés pour cette occasion. Les résultats ont été inscrits à la craie sur un tableau accroché à l'extérieur qui a été partiellement effacé par la pluie.

Classes (en cm)	**Nombre de Sapins**	**Fréquence**
[40-70[	**70**	**17,5%**
[70-100[		
[100-130[	**90**	
[130-160[		**25,0%**
[160-190[	**60**	

1/ Retrouver le comptage et le calcul des fréquences qui ont été effacées.

2/ Calculer les effectifs et les fréquences cumulés.

Exercice 7 :

Dans une entreprise de 70 personnes, l'ancienneté (en nombre d'année) des salariés se décompose comme suit :

Classes (en nombre d'année)	Nombre de salariés
[0-5[	**15**
[5-10[	**23**
[10-15[	**14**
[15-20[	**9**
[20-25[	**7**
[25-30[	**2**

1. Construire le tableau de distribution des effectifs, des fréquences, des fréquences cumulées, des effectifs cumulés croissants et des effectifs cumulés décroissants.
2. Combien de personnes ont plus de 10 ans d'ancienneté ?
3. Calculer l'ancienneté moyenne de l'entreprise.

Exercice 8 :

La répartition des jeunes âgés de **16** à **25** ans sans diplômes et résidant à Sarh en 2010 selon leur sexe et leur activité est la suivante :

Type d'activité	***Hommes***	***Femmes***	***Total***
Actifs ayant un emploi	6 639	2 446	9 085
Chômeurs	2 544	1 850	4 394
Militaires du contingent	112	2	114
Inactifs	1 201	1 872	3 073
Total	10 496	6 170	16 666

1. Quelle est la taille de la population âgée de 16 à 25 ans résidant à Sarh en 2010 ?
2. Quelle est le type de la variable "Type d'activité" ? Quel est son mode ?
3. Quelle est la distribution de proportion (3 chiffres après la virgule) parmi la population des femmes, parmi la population des hommes, parmi la population totale ?
4. Quelle est la proportion des jeunes actifs dans cette population ?

Exercice 9:

La distribution des salaires horaires, en euros, des **N** employés d'une grande entreprise est donnée par :

Classes	Effectifs
[50 ; 100[	10
[100 ; 150[	14
[150 ; 200[	16
[200; 250[	**n**

Ces données sont incomplètes car, à la suite d'un accident, l'effectif de la dernière classe est illisible ; alors, on a décidé de le noter provisoirement par **n**. mais, on sait que la **médiane** de cette série statistique est de **153,125 € (euro).**

1. Exprimer la moyenne arithmétique de cette distribution en fonction de **n.**
2. Exprimer la médiane de cette série statistique en fonction de **n**, sachant que la classe médiane retenue est **[150 ; 200[**. Déterminer l'effectif **n** et puis le nombre d'employé **N** de cette entreprise.
3. Retrouver la valeur numérique de la moyenne arithmétique.