CAPACITES D'ORTHOPHONIE ET D'ORTHOPTIE

Statistiques descriptives

Gérard Soula
LERTIM
Faculté de Médecine – Marseille

Support de cours d'après le polycopié : Biostatistique en PCEM1
H. Chaudet, B. Giusiano, J. Gouvernet, M. Roux, E. Sanchez, M. Fieschi

Introduction générale

aux statistiques

Avant propos

- Objectif = initiation méthodologique à l'analyse et au traitement des données en biologie et en médecine.

Généralités

- La méthode statistique a pour but de:
    - dégager certaines propriétés d'un ensemble de mesures (ou d'observations)
    - ou de décrire cet ensemble (appelé population).

Population, échantillon

☞ Une population peut être tout aussi bien un groupe d'êtres humains ou un ensemble d'objets.
Tous ces éléments ont en commun un attribut ou encore une propriété qui caractérise cet ensemble d'éléments.

Exemple: Les individus de sexe masculin.

☞ Le statisticien n'étudie pas le caractère sur l'ensemble de la population, mais sur un échantillon extrait de la population:

- La taille de la population peut être très importante et le coût de l'enquête serait trop important.
- L'accès à tous les individus de la population est matériellement impossible.
- L'étude du caractère peut détruire les éléments de la population.
- Le nombre d'éléments constituant l'échantillon est appelé l'effectif ou la taille de l'échantillon.

☞ Un bon échantillon doit constituer une image réduite de l'ensemble de la population dont on veut étudier un caractère bien défini. Dans le cas contraire, on dit que l'échantillon est biaisé.

☞ Le choix de l'échantillon, le recueil des données nécessaires à l'étude que l'on se propose, constituent la partie fondamentale, la plus longue, de l'étude.

☞ Afin de généraliser les résultats obtenus sur l'échantillon, on désire que celui-ci représente le mieux possible la population cible c'est à dire celle sur laquelle porte l'étude.

Echantillonnage

☞ Comment choisir un échantillon pour qu'il soit représentatif?

Tirage au hasard

☞ Un échantillon ne doit en aucun cas être choisi par commodité. Afin de disposer d'un échantillon représentatif, il faut le constituer d'une manière aléatoire:

- un véritable tirage au sort ,
- utilisation des tables de nombres aléatoires

Stratification

☞ On subdivise la population en sous groupes (ou strates). On choisit ensuite l'échantillon en tirant au sort dans les strates. Chaque strate peut être représentée en fonction de son importance dans la population.

Exemples:
Si l'on veut faire une enquête épidémiologique sur l'hypertension artérielle, on pourra constituer un échantillon qui sera un modèle réduit de la population étudiée. En stratifiant de telle sorte qu'il respecte les mêmes proportions que la composition de la population quant aux catégories socioprofessionnelles, aux tranches d'âges, au sexe....

Dans un essai thérapeutique d'un traitement anticancéreux, on pourra définir les strates en tenant compte des facteurs pronostiques tels que: taille de la tumeur, extension locorégionale, métastase à distance,....

☞ Le prélèvement de l'échantillon doit être fait au hasard.

☞ Il n'est pas toujours facile de prélever un bon échantillon.

Exemple de difficultés dans le chois d'un échantillon:
On se propose d'étudier le pourcentage de décès dans la population française des sujets atteints d'un infarctus du myocarde.
On peut constituer un échantillon en observant les décès des malades qui ont été hospitalisés dans un service hospitalier donné. Le biais introduit, si la population “cible” est la population de tous les français, est évident.
En effet, le service hospitalier a un recrutement particulier et une renommée telle qu'il hérite, peut-être, de malades plus graves, ou d'une catégorie sociale dont le genre de vie, l'alimentation, l'âge...., sont des facteurs pronostiques qui peuvent modifier l'issue de la phase aiguë.
Un échantillon représentatif de la population française atteinte d'un infarctus du myocarde pourrait être obtenu par tirage au sur tous les cas d'infarctus du myocarde recensés en France. Toutefois on ne les connaît pas tous et il est toujours possible d'introduire un biais.

Problème de l’estimation

☞ Il s’agit d’évaluer un paramètre (une caractéristique) sur un échantillon pour pouvoir estimer ce paramètre pour la population toute entière.

Exemple:
Evaluation, à partir de la mesure de la glycémie pratiquée sur un échantillon de sujets sains ayant entre 20 et 40 ans, de la valeur moyenne de la glycémie pour tous les sujets sains de cette tranche d'âge.
Si l'on veut que cette estimation soit aussi précise que possible, il est nécessaire que l'échantillon soit aussi représentatif que possible de la population.

Les tests statistiques

☞ Problème: de tirer des conclusions sur la population à partir de l'étude d'un ou plusieurs caractères observés sur les individus d'un ou de plusieurs échantillons issus de cette population.
Ce problème inclut celui de la comparaison de caractéristiques (une ou plusieurs) issues de 2 ou plusieurs populations. (comparer la glycémie moyenne des sujets urbains et des sujets ruraux).

☞ Solution: les tests statistiques qui sont des tests d'hypothèses. Ils permettent de faire des inférences statistiques.

Statistique descriptive

Buts de la statistique descriptive

☞ Toute série d'observations comporte un certain nombre de données relatives à un ou plusieurs caractères ou encore variables.

☞ Le but des statistiques descriptives est de décrire un ensemble d'observations à l'aide de quelques éléments caractéristiques.

Exemple: la taille des Français adultes.
Dans ce cas les mesures seront nombreuses, le tableau des données, c'est-à-dire la liste des tailles de tous les sujets, ne donnera, au premier abord, aucun renseignement clair. Grâce aux statistiques descriptives, on caractérisera cet ensemble d'individus par un moyen simple qui permettra de réduire le nombre de données. Par exemple, si on s'intéresse à la taille, on procédera au calcul de la valeur moyenne de la taille.

☞ De cette façon, il est bien certain que l'on perd de l'information, mais on gagne en commodité de manipulation.

☞ Pour présenter les données, le premier travail consiste donc à rassembler et à présenter clairement les observations. Plusieurs cas sont à envisager suivant le type des données recueillies: qualitatif, ordinal, quantitatif.

Les différents types de données

1) Les données de type qualitatif

☞ Un caractère est qualitatif s'il peut se présenter sous plusieurs aspects ou suivant plusieurs modalités.

☞ Ces données donnent lieu à des dénombrements.

Exemples: le sexe, la couleur des yeux, l'efficacité ou la non efficacité d'un traitement, la nature des cellules d'un tissu, le groupe sanguin,....

☞ On est amené à définir des catégories ou classes exclusives correspondant aux différentes modalités du caractère observé, puis à déterminer à quelle classe appartient chaque individu.

☞ Un individu appartient à une classe et une seule.

2) Données de type ordinal

☞ Le caractère observé est de type ordinal s'il existe entre les diverses classes une relation d'ordre.

Exemple de relation d'ordre: plus grave que...., de meilleur pronostic que....

Exemple: Classification en stades 1, 2, 3, 4 des patients atteints de la maladie de Hodgkin.
Les malade au stade 2 sont plus gravement atteints que ceux qui sont classés au stade 1...

☞ on affecte chaque individu à une classe et une seule.

☞ Il existe un ordre sur les classes.

3) Données de type quantitatif

☞ Une variable quantitative prend pour valeur un nombre résultant de la mesure, avec une unité, du caractère chez chaque individu.

☞ La mesure est telle qu'une même différence entre des valeurs observées a toujours la même signification.

Exemple de la mesure de la taille:
Soient 4 individus A, B, C, D dont les tailles sont exprimées en centimètres: A = 175 cm; B = 180 cm; C = 165 cm; D = 170 cm.
(On peut dire que 180 - 175 = 170 - 165 = 5 cm.)

☞ Un caractère quantitatif est :

- discret s'il prend des valeurs isolées,
- continu s'il prend toutes les valeurs,

de son intervalle de variation.

Exemples de caractères quantitatifs discontinus (ou discrets):
nombre d'enfants dans une fratrie, nombre de cellules par mm^3,...

Exemples de caractères quantitatifs continus: tension artérielle, glycémie,...)

Caractérisation des données

I) Caractérisation des données qualitatives et ordinales unidimensionnelles

II) Caractérisation des données qualitatives à deux dimensions

III) Caractérisation des données quantitatives à une dimension

IV) Caractérisation des données quantitatives à deux dimensions

Caractérisation des données qualitatives et ordinales unidimensionnelles

1) Fréquence absolue et tableau des effectifs

2) Fréquences relatives

3) Fréquences cumulées (relatives et absolues)

4) Diagramme "camembert"

5) Diagramme en bâtons et mode

1) Fréquence absolue et tableau des effectifs

☞ La fréquence absolue est le nombre d'individus par classe.

☞ Ce dénombrement donne lieu à une représentation des données sous forme de tableau.

Exemple: On a dénombré sur une ensemble de 180 sujets, les individus qui appartenaient aux différents groupes sanguins.

A+	A-	B+	B-	AB+	AB-	O+	O-
80	10	20	5	5	2	50	8

Description de l'échantillon des groupes sanguins

☞ Sur les classes ainsi formées, seules les opérations suivantes sont permises:

- réaliser des classes disjointes à partir d'une seule classe,
- regrouper certaines classes.

☞ La seule relation qui puisse être utilisée sur ces données est la relation d'appartenance à une même classe.

Exemple: regrouper les classes correspondant aux rhésus + ou -, ou ignorer le rhésus pour former les groupes A, B, AB, O

A	B	AB	O
90	25	7	58

Description de l'échantillon des groupes sanguins
sans facteur rhésus

2) Fréquences relatives

☞ Les fréquences relatives sont, pour chaque classe, le rapport de son effectif au nombre total d'individus de la série des mesures.

☞ La somme des fréquences relatives est égale à 1.

☞ Parfois, les résultats sont exprimés en pourcentage, chacune des fréquences relatives étant multipliée par 100 et arrondies à l'unité

A	**B**	**AB**	**O**
50%	14%	4%	32%

Fréquences relatives (exprimées en pourcentage)

3) Fréquences cumulées (relatives et absolues)

☞ Les fréquences cumulées sont utilisées pour les données ordinales qui présentent des classes ordonnées.

Exemple: Sur un échantillon de 500 malades cancéreux, on a noté le stade de la maladie. On peut résumer ou présenter ces données par des fréquences relatives.

Stade	**Nombre de malades**	**Fréquence relative (%)**	**Fréquence relative cumulée(%)**
1	350	70	70
2	110	22	92
3	30	6	98
4	10	2	100

Répartition du Stade de la maladie

Cette présentation permet de dire, par exemple, que 92% des sujets examinés ont un stade inférieur ou égal à 2.

4) Diagramme "camembert"

☞ On peut représenter les effectifs absolus ou relatifs des classes par des secteurs de cercle dont la surface est proportionnelle à l'effectif.

☞ Le diagramme "camembert" est bien adapté à la représentation des données qualitatives "pures".

Exemple:

Yeux	**marron**	**vert**	**bleu**	**noir**
Effectif	50	10	28	12

Couleur des yeux dans un échantillon de 100 sujets

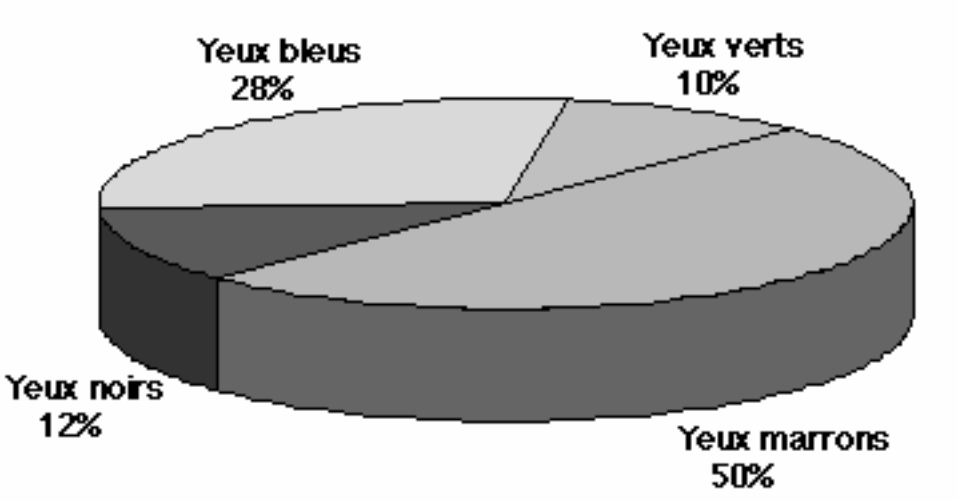


Diagramme "camembert"

5) Diagramme en bâtons et mode

5.1) Diagramme en bâtons

Pour les données ordinales on peut également représenter les fréquences absolues, relatives ou cumulées par un diagramme en bâtons.

Exemple: échantillon de 500 cancéreux dont on a noté le stade.

Stade	**Nombre de malades**	**Fréquence relative (%)**	**Fréquence relative cumulée**
1	350	70	70
2	110	22	92
3	30	6	98
4	10	2	100

Répartition du Stade de la maladie

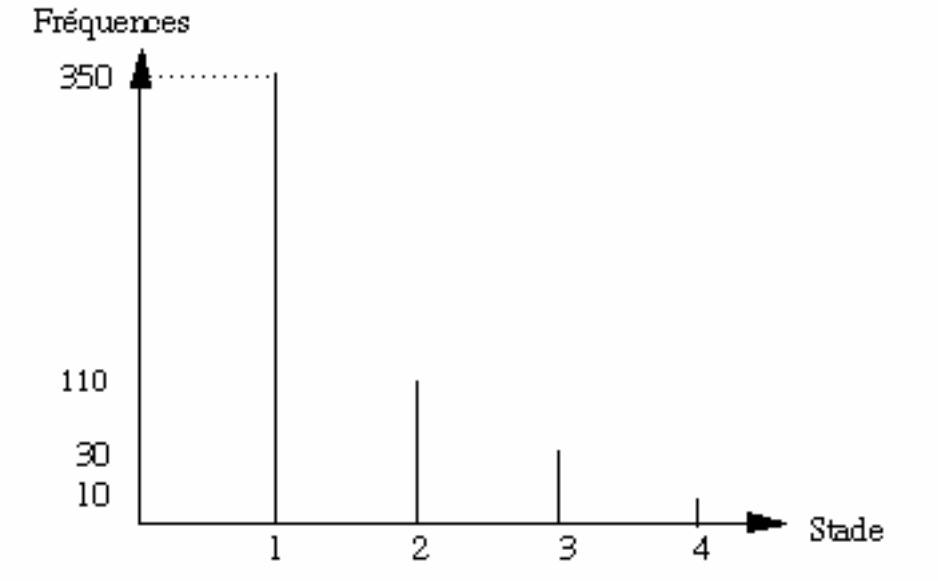


Diagramme en bâtons des stades de la maladie

5.2) Mode

☞ Le mode ou classe modale est la classe (catégorie) qui offre la plus grande fréquence.

Exemple:

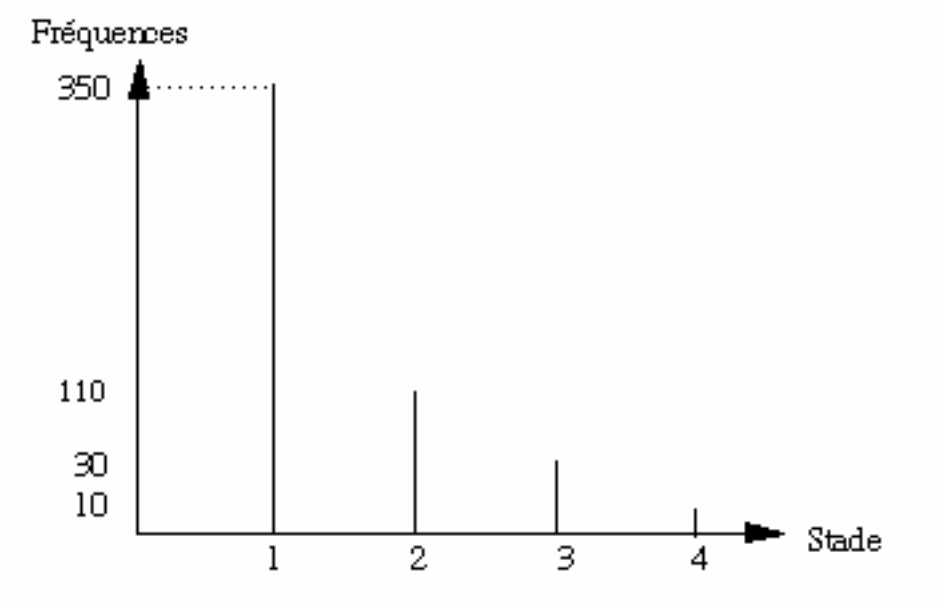


Diagramme en bâtons des stades de la maladie

La classe caractérisée par le stade 1 est la classe qui contient le plus grand nombre de sujets, c'est le mode ou classe modale.

☞ Dans le cas de variables ordinales, si les données montrent plusieurs classes d'effectifs supérieurs aux effectifs des classes voisines, on dit que le diagramme représente une distribution multimodale: bi-modale, tri-modale... Dans le cas contraire, on dit que la distribution est uni-modale.

Caractérisation des données qualitatives

à deux dimensions

☞ Les modalités de deux variables qualitatives permettent de constituer des classes exclusives auxquelles sont affectées chaque observation.

☞ Les classes exclusives sont représentées sous la forme d'un tableau appelé tableau de contingence.

Exemple: Dans un échantillon de 200 sujets on a relevé la présence ou l'absence d'un signe clinique S cet d'une maladie M.
Les malades présentant la maladie sont dénombrés dans la colonne M, les autres dans la colonne non M.

	M	**non M**	**Total**
S	90	30	120
non S	30	50	80
Total	120	80	200

Tableau de contingence

☞ Ce tableau comporte deux parties:

- les effectifs dénombrés pour chacune des modalités, pour chacun des deux caractères étudiés,
- les effectifs de chaque modalité d'un caractère, quelles que soient les modalités de l'autre caractère. Ces effectifs sont situés dans la dernière colonne et la dernière ligne.

La dernière ligne et la dernière colonne sont appelées: les "marginales", (marge ligne et marge colonne) ou encore "distributions marginales".

Caractérisation des données quantitatives à une dimension

1) Généralités

☞ Rappel: les variables quantitatives peuvent être de deux types: variables discontinues (ou discrètes) et variables continues.

☞ Dans le cas des variables discontinues, il est possible de représenter les données par un diagramme en bâtons, comme dans le cas de données ordinales.

☞ Dans tous les cas, on peut diviser l'intervalle de variation de la variable en un certain nombre de classe et l'on dénombre toutes les mesures à l'intérieur de chaque classe.

Exemple: soit la série de mesures représentant les âges de 20 individus, rangées par ordre croissant:
3, 5, 6, 7, 8, 11, 15, 20, 21, 22, 23, 23, 23, 30, 31, 32, 35, 36, 40, 45
On peut décider de déterminer des classes d'âge de 10 ans en 10 ans:
0-10 ans, 10-20 ans, 20-30 ans, 30-40 ans, 40-50 ans.

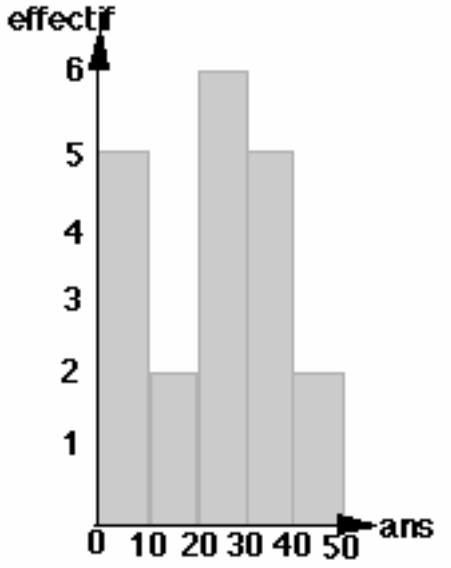
Classe	Effectif/Classe
1	5
2	2
3	6
4	5
5	2

Effectifs par classe

2) Histogramme

☞ Construction: on porte sur l'axe des abscisses les extrémités de chaque classe et pour chacune d'elles on construit un rectangle dont la base est le segment limité aux extrémités de la classe et la surface est proportionnelle à l'effectif de la classe.

☞ La surface limitée par la ligne polygonale obtenue en bordant la partie supérieure de l'ensemble des rectangles s'appelle l'histogramme.



Histogramme

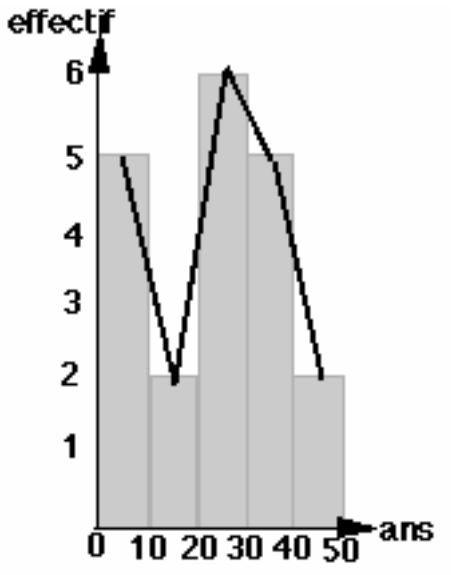
☞ Un histogramme est tracé en respectant deux règles:

- l'échelle sur l'axe des abscisses est identique pour tous les intervalles de classes,
- la surface de chacun des rectangles est proportionnelle au nombre d'individus de la classe.

La deuxième règle se simplifie si les intervalles de classe ont tous la même largeur. Cette simplification est très souvent utilisée. En effet quand les intervalles de classe sont de même largeur, la hauteur du rectangle est proportionnelle à l'effectif, ce qui facilite la lecture de

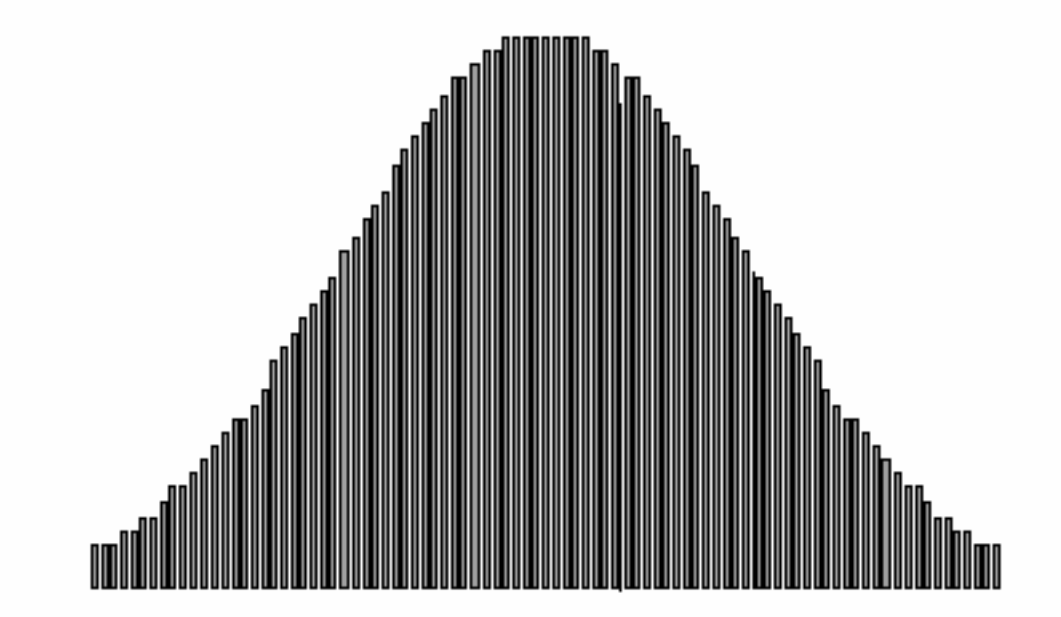
l'histogramme.

☞ Le contour polygonal joignant les milieux des bases supérieures des rectangles s'appèle le polygone des fréquences.



Histogramme et polygone des fréquences

☞ Si on augmente le nombre des classes recouvrant l'étendue de l'échantillon, l'intervalle de chaque classe devenant très petit, on peut admettre, à condition que la population soit "infinie", que l'histogramme et le polygone des fréquences se "rapprochent", et que leur limite commune est une courbe continue.



Courbe de fréquences et distribution des fréquences

Cette courbe est dite "courbe des fréquences".

☞ L'ensemble des classes affectées de leur fréquence constitue une distribution de fréquences.

☞ Quand le nombre de classes tend vers l'infini, le polygone des fréquences devient une ligne continue: la courbe des fréquences.

3) Paramètres statistiques décrivant un ensemble de mesures quantitatives.

☞ Problème: présenter de façon simple et abrégée, les caractéristiques principales de l'ensemble des mesures qui ont été effectuées sur un échantillon ou une population.

☞ Solution: utiliser des grandeurs numériques appelées paramètres de la distribution réparties en deux catégories:

- les paramètres de position: moyenne, médiane, mode
- les paramètres de dispersion: variance, écart-type, quantiles.

☞ Ces paramètres font partie des grandeurs statistiques que l'on nomme parfois "statistiques". On dira par exemple que le calcul de la moyenne est le calcul d'une statistique.

3.1) Paramètres de tendance centrale ou de position: moyenne, médiane, mode

☞ Ces paramètres définissent l'ordre de grandeur de l'ensemble des mesures de la distribution. Ce sont les valeurs autour desquelles se groupent les différentes mesures.

La moyenne

☞ La valeur centrale la plus utilisée est la moyenne arithmétique de la variable mesurée, c'est-à-dire le rapport de la somme des mesures au nombre de mesures effectuées.

On peut caractériser la tendance centrale par la moyenne notée:

$$\overline{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \left(\frac{x_i}{n} \right)$$

☞ La moyenne s'exprime dans les mêmes unités que les valeurs observées.

☞ Notations:

- $\{x_1, x_2, \ldots\ldots x_n\}$ la série de mesures,
- $\overline{X}$ la moyenne,
- n l'effectif.

Exemple:
Considérons la série de mesures constituée par les poids en Kg de 5 individus: 72,5 ; 70,0 ; 76,0; 68,5 ; 73,0.
La moyenne est égale à 72 kg.

La médiane

☞ La médiane est l'abscisse du point qui laisse de part et d'autre un nombre égal d'observations. C'est donc un nombre de même nature et de même unité que les valeurs observées.

☞ Pour déterminer la médiane d'une série de nombres, il est nécessaire d'ordonner cette série de mesures.

Exemple :
poids en Kg de 5 individus: 72,5 ; 70,0 ; 76,0; 68,5 ; 73,0.

mesures ordonnées: 68,5 ; 70 ; 72,5 ; 73,0 ; 76,0.

La médiane est égale à 72,5 Kg.
(il y a autant de mesures inférieures à 72,5 que de mesures supérieurs à72,5) .

☞ Deux cas peuvent se présenter:

- si n est impair (n = 2k + 1), la médiane est la valeur de la mesure qui se situe au milieu de la série de mesures ordonnées: c'est x_{k+1}

- Si n est pair (n = 2k), on appelle médiane toute valeur comprise entre x_k et x_{k+1}. En effet, il n'y a pas de valeur observée qui soit au milieu de la série de mesures.

Remarques:

- La médiane est moins influencée que la moyenne arithmétique par les valeurs extrêmes de la variable.

  Exemple :Si le premier des poids (68,5 kg) est remplacé par 55 kg, la moyenne est influencée alors que la médiane reste identique.

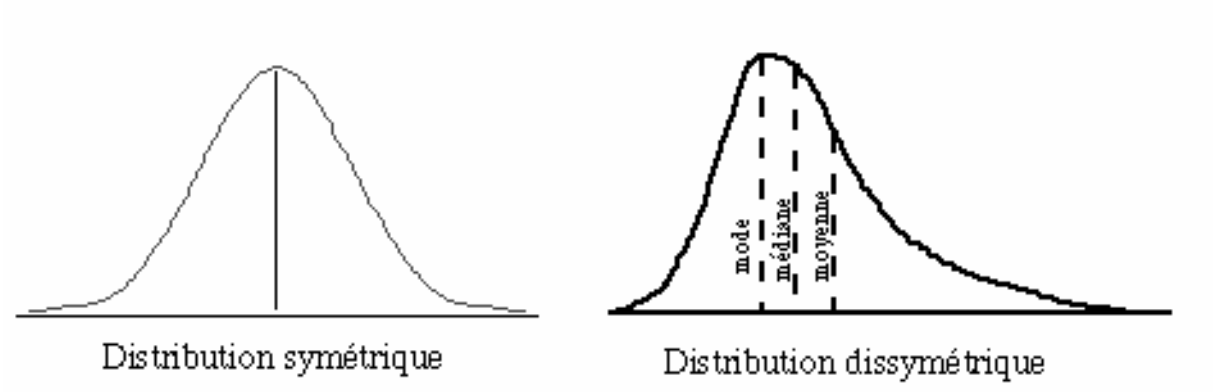
- La médiane peut aussi être utilisée dans le cas des données ordinales, puisque sa détermination se base sur l'ordre des données.

Le mode

☞ Le mode, encore appelé valeur dominante, est la valeur de la variable dont la fréquence est maximale.

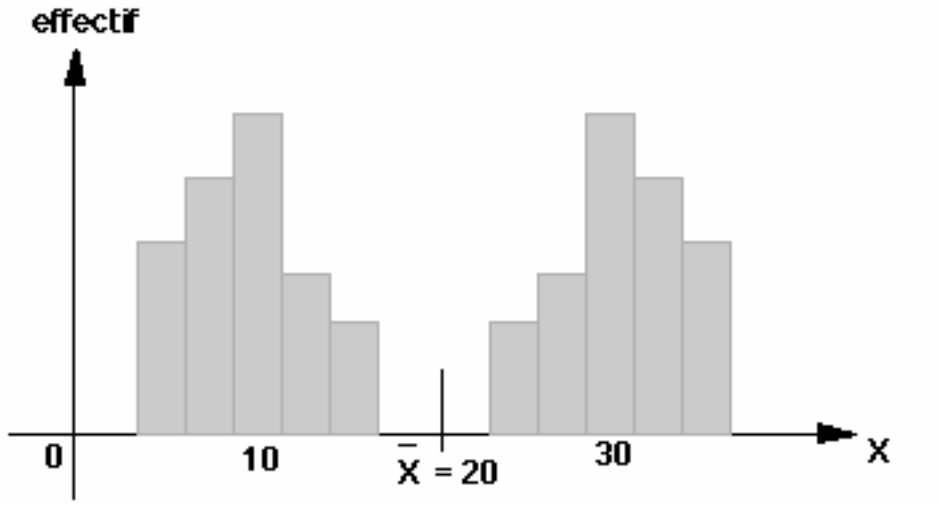
☞ Si les données sont affectées à des classes, on parle de classe modale. La classe modale est celle dont la fréquence est maximale. C'est, dans le cas d'une courbe continue des fréquences, l'abscisse du point d'ordonnée maximale.

☞ Si la distribution de fréquences est symétrique et unimodale, moyenne arithmétique, médiane et mode sont confondus.



Distribution symétrique et dissymétrique

☞ Il peut, dans certaines distributions, n'y avoir qu'un petit nombre d'observations dans le voisinage de la moyenne ou de la médiane Aucune valeur de X ne se trouve près de la moyenne et de la médiane.



Moyenne et dispersion

☞ Une distribution peut avoir plusieurs classes modales, elle est dite alors: bimodale, trimodale,....., plurimodale (plusieurs classes dont les effectifs sont grands, séparées par des classe à effectifs faibles). Une telle distribution traduit généralement un échantillon d'individus hétérogènes.

☞ Propriété de la moyenne

Soit la série de mesures $\{x_1, x_2,....x_n\}$ de moyenne $\bar{X}$ et deux constantes a et b ($a \neq 0$).

Considérons la série: $\{y_1,..... y_n\}$
telle que $y_1 = ax_1 + b$, $y_2 = ax_2 + b$,..... $y_n = ax_n + b$;
soit $\{ax_1 + b;....; ax_n + b\}$

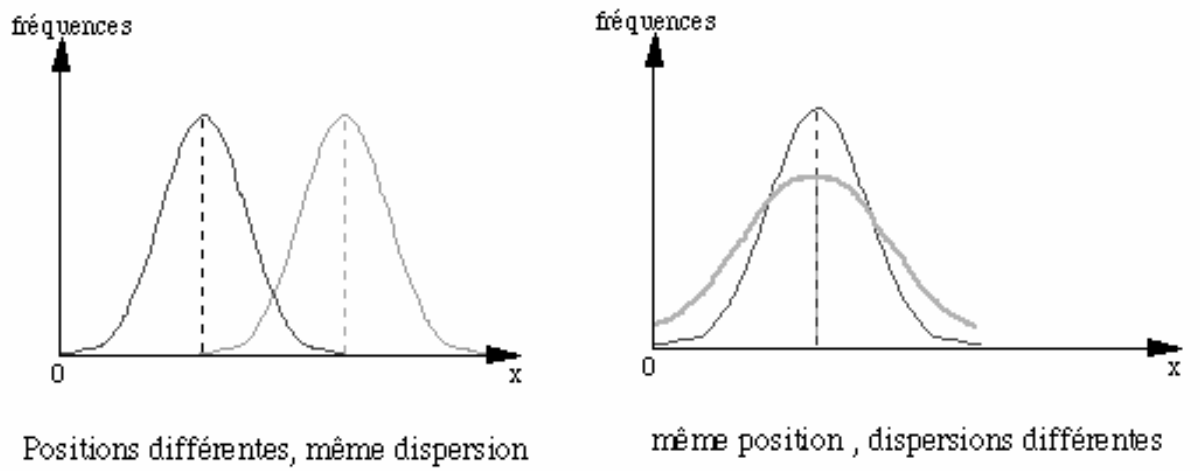
La moyenne de cette nouvelle série est égale à $\bar{Y} = a\bar{X} + b$

3.2) Paramètres de dispersion

☞ Problème: la moyenne ne suffit pas pour caractériser un ensemble de données.

Exemples:

- *La valeur moyenne de la série suivante: 1, 8, 9, 10, 11, 12, 19 est égale à 10.*
- *La valeur moyenne de la série 8, 8, 9, 10, 11, 12, 12 est égale à 10. La dispersion des mesures autour de la moyenne 10 est beaucoup moins importante.*



Paramètres de dispersion et de position

☞ Solutions: Variance, écart-type, Etendue, Quantiles

Variance

☞ Le paramètre le plus efficace pour rendre compte de la dispersion d'une série de mesures est la variance ou sa racine carrée: l'écart-type.

☞ La variance est définie comme la moyenne arithmétique des carrés des écarts à la moyenne de l'échantillon.

$$Var(X) = \sum_{i=1}^{n} \frac{(x_i - \overline{X})^2}{n}$$

où $\overline{x}$ est la moyenne de la série de mesures et n l'effectif.

☞ Propriété de la variance

$$Var(X) = \frac{\sum_{i=1}^{n} x_i^2}{n} - \overline{x}^2$$

"Moyenne des carrés moins carré de la moyenne"

☞ Attention: Var(X) est la variance de l'échantillon, ce n'est ni la variance de la population dont est issu l'échantillon, ni l'estimation de la variance de la population.

Ecart-type

☞ l'écart-type est la racine carrée de la variance, afin de disposer d'un indice de dispersion qui s'exprime dans la même unité que la grandeur mesurée.

$$Ecart-type(X)=\sqrt{Var(X)}$$

Exemple:
calcul de variance et d'écart-type de la mesure des poids d'individus dans un échantillon de moyenne 72 kg.

x_i	70	68,5	72,5	73	76
$x_i - \bar{X}$	-2	-3,5	0,5	1	4
$(x_i - \bar{X})^2$	4	12,25	0,25	1	16

Variance = 33,5 / 5 = 6,7 Kg^2

Ecart-type = 2,59Kg.

x_i	70	**70**	72,5	73	76

Ecart-type = 2,23Kg.

x_i	70	**60**	72,5	73	76

Ecart-type = 5,5Kg.

☞ Propriété de la variance:

Soit la série $\{x_1, x_2,.....x_n\}$ de moyenne $\bar{x}$ et de variance Var (X) et deux constantes a et b ($a \neq 0$).

Considérons la série $\{y_1,..., y_i,...,y_n\}$ sachant que $y_i = ax_i + b$.
Soit $\{y_1 = ax_1 + b;.....;y_n = ax_n + b\}$.

Le calcul montre que la variance de cette nouvelle série (que nous noterons Var(Y)) est égale à:

$$Var(Y) = a^2 Var(X)$$

$$ecart-type(Y) = |a| \cdot \sqrt{Var(x)} = |a| \cdot ecart-type(X)$$

Etendue

☞ On définit l'étendue d'une série de mesures comme la différence entre la plus grande et la plus petite valeur de la série:

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

Quantiles

☞ Les quantiles sont les valeurs de la variable qui divisent l'échantillon ordonné en groupes d'effectifs égaux.

☞ Selon le nombre de groupes souhaités les quantiles portent des noms différents:

- quartiles pour 4 groupes,
- déciles pour 10 groupes,
- percentiles pour cent groupes.

Les quartiles

☞ Les quartiles sont les abscisses correspondant aux valeurs de la série X qui partagent l'effectif total, après l'avoir ordonné, en 4 classes de même effectif.

☞ Entre la limite inférieure de l'étendue et le premier quartile, on retrouve un quart des observations. Un autre quart des observations se retrouve entre le premier et le deuxième quartile.

☞ Remarque: le deuxième quartile n'est autre que la médiane.

Exemple:
Ages de n = 20 individus: 3, 5, 6, 7, 8, 11, 15, 20, 21, 22, 23, 23, 23, 30, 31, 32, 35, 36, 40, 45.

L'effectif de chaque quartile est de 5.

La valeur qui sépare le 1er groupe du 2ème est située entre 8 et 11.

Toute valeur comprise entre 8 et 11 peut être retenue comme premier quartile, toute valeur entre 22 et 23 comme deuxième quartile.

☞ Les écarts inter-quartiles donnent une idée de la dispersion: plus les données sont groupées, plus l'espace inter-quartile est faible. (Différence entre la première et dernière valeur contenue dans le quartile).□

Les percentiles

☞ Les percentiles définissent 100 groupes d'effectifs correspondants chacun à 1% de l'effectif de l'échantillon.

☞ Le 50^{ième} percentile correspond à la médiane.

☞ Remarque: Mode, médiane, quantiles peuvent être utilisés dans le cas de données ordinales

Caractérisation des données quantitatives à deux dimensions

☞ Considérons des mesures X et Y effectuées sur un échantillon d'effectif n. Deux mesures ont été effectuées en même temps sur le même individu "i": x_i et y_i.

Exemple: On peut mesurer chez n individus la concentration sanguine du potassium et du chlore. Nous obtenons pour chaque individu le couple de mesures(x_i , y_i).

☞ Problème: comment caractériser ces mesures ?

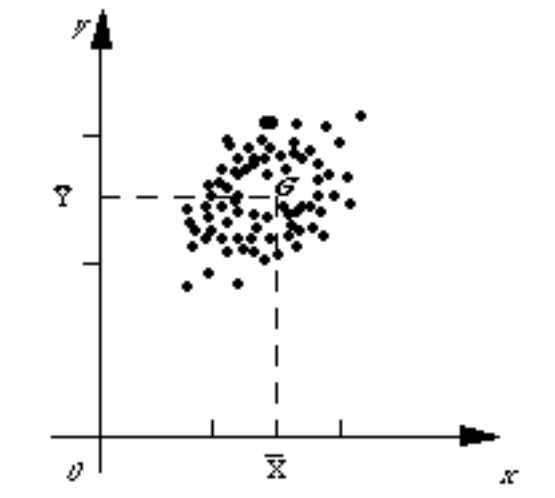
☞ Solutions: nuage de points, covariance, coefficient de corrélation

1) Nuage de points

☞ Chacune des observations (x_i , y_i) est représentée par un point dans le plan.

☞ L'ensemble des points constitue "un nuage de points". L'aspect de ce nuage est important à observer.

☞ Le point ayant pour coordonnées ($\bar{X}$,$\bar{Y}$) représente le point "moyen". Ce point, noté G, dont les coordonnées sont les moyennes des deux séries d'observations, s'appelle: <u>"centre de gravité"</u>.



<u>Nuage de points</u>

2) Covariance

☞ Rappel: Variance = dispersion d'une donnée quantitative.

☞ Covariance = dispersion d'une donnée quantitative à deux dimensions.

☞ La covariance caractérise la dispersion des points de coordonnées (x_i , y_i) par la moyenne du produit des écarts, pour chaque point, entre ses coordonnées et leurs valeurs moyennes.

$$Covar(X,Y)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{X})\cdot(y_i-\overline{Y})$$

3) Coefficient de corrélation

☞ Le coefficient de corrélation, noté r, est une grandeur qui reflète la dispersion des couples (x_i , y_i) en fonction de la dispersion observée pour chacune des deux séries de mesures X et Y.

$$r=\frac{Covar(X,Y)}{\sqrt{Var(X)\cdot Var(Y)}}$$

Remarque: r est un nombre sans dimension et on montre qu'il est compris entre -1 et 1.

Ce qu'il faut savoir absolument

Type de données	Présentation des données	Représentation graphique	Données à 2 dimensions
Qualitatif	Dénombrement par classes Fréquences absolues Fréquences relatives	Camembert	Tableau de contingence Regroupement libre
Ordinal	Dénombrement par classes ordonnées Fréquences absolues Fréquences relatives Fréquences cumulées	Diagramme en bâton	Tableau de contingence Regroupement des classes contiguës
Quantitatif	Tableau des mesures Paramètres de tendance centrale (ex: moyenne) Paramètres de dispersion (ex: variance, écart-type)	Histogramme	Nuage de points Covariance Coefficient de corrélation

Représentation des différents types de données

Type de données	Paramètre central	Calcul	Unité
Quantitatif	moyenne	$\overline{X}=\frac{1}{n}\cdot\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}\left(\frac{x_i}{n}\right)$	celle des x_i
	médiane	Valeur(s) qui sépare(nt) l'ensemble des données ordonné(es) en deux sous ensembles de même effectif	celle des x_i
Ordinal	mode	classe de plus grand effectif	celle des x_i

Paramètres de tendance centrale d'une série d'observations

Type de données	Paramètre de dispersion	Calcul	Unité
Quantitatif	variance	$Var(X)=\sum_{i=1}^{n}\frac{(x_i-\overline{X})^2}{n}$	unité des x_i au carré
	écart-type	$\sqrt{Var(X)}$	celle des x_i
	quantiles	valeur de la variable divisant l'ensemble ordonné des données en sous ensembles d'effectifs égaux	celle des x_i
Ordinal	quantiles	valeur de la variable divisant l'ensemble ordonné des données en sous ensembles d'effectifs égaux	celle des x_i

Paramètres de dispersion