Université de Blida 1
Faculté des Sciences de la nature et de la vie
Département d'Agro-alimentaire

Matière : Traitement des données d'analyse

M2 SAACQ

Presenté par Dr. HAMIDI
youcefhamidi982@yahoo.fr

1.1 Définitions fondamentales

La **statistique** est une méthode scientifique qui consiste à réunir des données sur des ensembles, puis à analyser, à commenter et à critiquer ces données pour évaluer la «fiabilité » des décisions fondées sur ces «données».

Elle est appliquée a la plupart des disciplines : agronomie, biologie, démographie, économie, sociologie, linguistique, psychologie, . . .

Pourquoi enseigner la statistique ou le traitement des données?

- Expert ?!!!
- Il faut être autonome
- Quand vous connaissez la statistique, vous n'expérimentez plus de la même manière

Les ensembles étudiés sont appelés **population**. Les éléments de la population sont appelés **individus** ou unités statistiques. La population est étudiée selon un ou plusieurs **caractères**.

Les **statistiques descriptives** peuvent se résumer par le schéma suivant :

Population :C'est l'ensemble des « individus» à propos desquels on souhaite pouvoir tenir des décisions. Elle est le plus souvent définie par une propriété portant une ou plusieurs variables :

- ✓ L'ensemble des nouveaux nés de mère diabétique;
- ✓ L'ensemble des hommes obèses.

2. Echantillonnage statistique

Pour recueillir des informations sur une population statistique, l'on dispose de deux méthodes :

- la **méthode exhaustive** ou recensement où chaque individu de la population est étudié selon le ou les caractères étudiés.

- la **méthode des sondages** ou échantillonnage qui conduit à n'examiner qu'une fraction de la population, un **échantillon**.

2.1 Définition

L'**échantillonnage** représente l'ensemble des opérations qui ont pour objet de prélever un certain nombre d'individus dans une population donnée.

Pour que les résultats observés lors d'une étude soient généralisables à la population statistique, **l'échantillon doit être représentatif** de cette dernière, c'est à dire qu'il doit refléter fidèlement sa composition et sa complexité. Seul **l' échantillon aléatoire** l'échantillonnage aléatoire assure la représentativité de l'échantillon.

Un échantillon est qualifié d'**aléatoire** lorsque chaque individu de la population a une **probabilité connue et non nulle** d'appartenir à l'échantillon.

Séance 2

2. Echantillonnage statistique

2.2 Echantillonnage aléatoire simple

L'**échantillonnage aléatoire simple** est une méthode qui consiste à prélever **au hasard** et de **façon indépendante**, ***n*** individus ou unités d'échantillonnage d'une population à ***N*** individus.

3. Les caractères statistiques

Le **caractère** désigne une grandeur ou un attribut, observable sur un individu et susceptible de **varier** prenant ainsi différents états appelés **modalités**.

Exemple :

Lors des recensements, les caractères étudiés sont l'âge, le sexe, la qualification professionnel, etc. Le caractère « sexe » présente deux modalités alors que pour la qualification professionnelle, le nombre de modalités va dépendre de la précision recherchée.

Donnée - Variable

Donnée (valeur) = résultat de l'observation d'un individu

- ✓Observer : réduire un objet infiniment complexe à un nombre limité de caractéristiques;
- ✓ Bien choisir la « caractéristique» reflète déjà une connaissance sur l'individu ;
- ✓ Observer : nécessite un instrument de mesure (au minimum, certains sens de l'observateur lui-même, souvent un appareil artificiel)

Types de variables

- Variable quantitative

Toute variable qu'un instrument peut mesurer sous forme numérique

✓ Taille, poids, glycémie, nombre d'enfants dans une fratrie, …

Les valeurs possibles pour une telle variable sont donc:

✓ l'ensemble des réels;

✓ Plus souvent un sous-ensemble de l'ensemble des réels

- Entiers (naturels ou relatifs)
- Intervalle [a , b]

Types de variables

- Variable qualitative

Toute variable caractérisée par un attribut qualitatif, et non par une mesure numérique

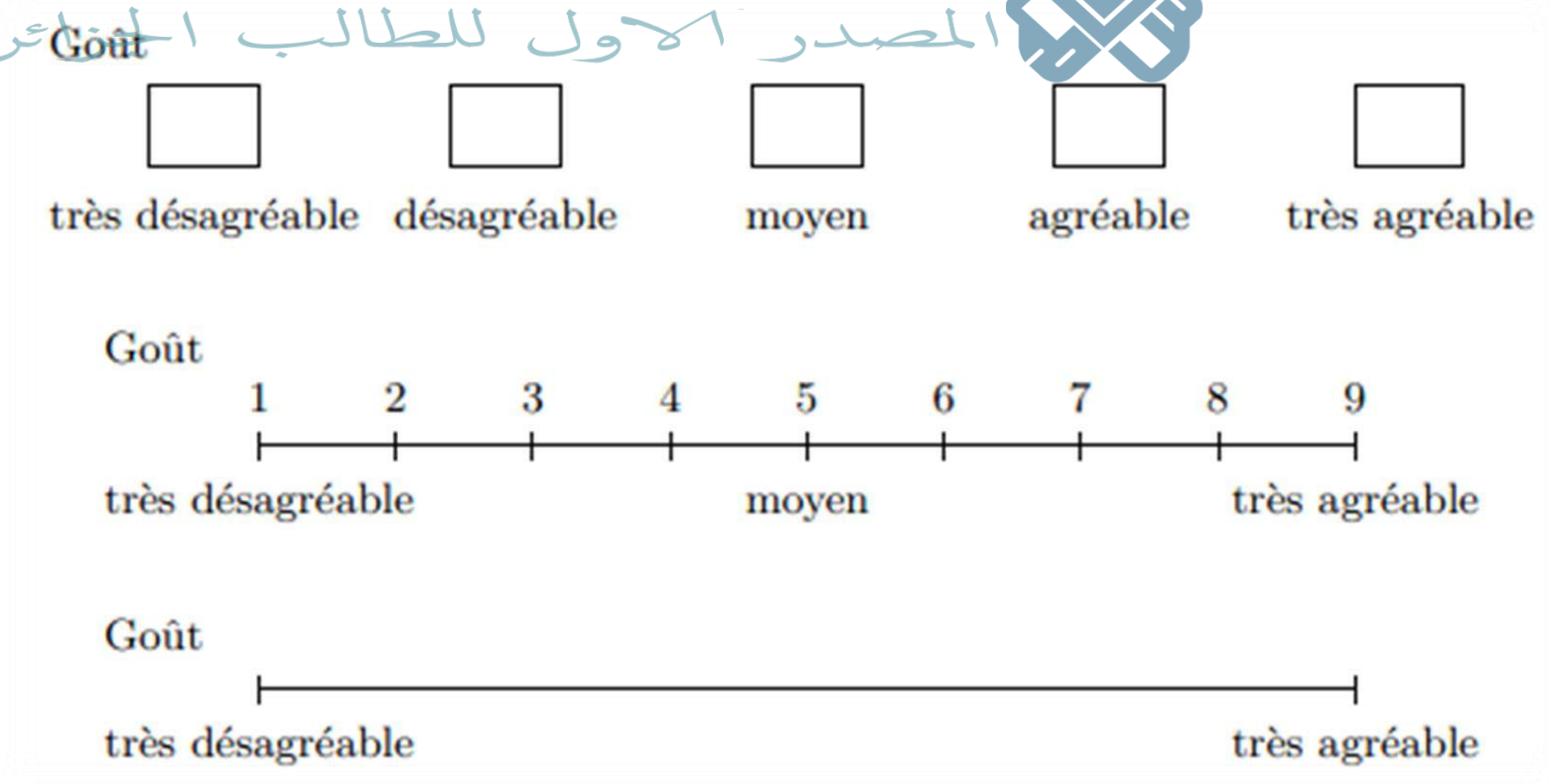
✓Couleur des yeux, sexe, présence d'un facteur de risque pour une pathologie, …

– Les valeurs possibles pour une telle variables, encore appelées les « modalités de réponse » sont donc:

✓ Une liste de modalités de réponse :

– Pour la couleur des yeux : {noir ; bleu ; vert}

– Pour le sexe : {homme ; femme}



Exemples d' échelles pouvant être utilisées dans une analyse sensorielle

Variable ordinale (ou pseudo-quantitative)

Toute variable qualitative dont les valeurs peuvent être «ordonnées » (on peut classer les valeurs possibles par ordre « croissant »)

✓ Intérêt d'un spectateur pour un film : {nul ; moyen ; fort; passionné } ou {0 ; 1 ; 2 ; 3} ou …

✓ Mention au baccalauréat : {ajourné ; passable ; assez bien ; bien; très bien} ou {AJ ; P ; AB ; B ; TB}ou {0 ; 1 ; 2 ; 3 ; 4} ou …

Deux grands sous-types de variables quantitatives :

- **Variable continue**: elle peut prendre n'importe quelle valeur dans un intervalle donné (à condition d'avoir un instrument de mesure suffisamment précis)
  - ✓ Taille, poids, glycémie
- **Variable discrète**: elle ne peut prendre qu'un nombre fini ou dénombrable de valeurs (on peut « compter» les valeurs possibles)
  - ✓ Nombre de globules blanc dans un volume de 1 ml, rapport entre le nombre d'ailes et le nombre de pattes d'un insecte, …

Paramètres

- Un paramètre est un grandeur apportant d'une information résumée sur le variable d'intérêt. Il est soit mesuré dans un échantillon, soit estimé dans la population, à partir des observations de l'échantillon

5.5. Observation

Observer : réduire un objet infiniment complexe à un nombre limité de caractéristiques. Ceci nécessite un instrument de mesure.

5.5.1. Les différents types d'observation

✓Des observations préliminaires ou initiales sont réalisées avant ou au début d'application des différentes traitements. Ces observations sont essentiellement destinées a pouvoir disposer d'un certains nombres d'informations relatives aux différentes unités expérimentales avant tout traitements.

✓ Les observations qui sont faites au termes de l'expérience et qui sont directement liées aux objectifs sont généralement considérés comme étant les observations principales. Elles sont relatives a une ou a plusieurs variables ou caractéristiques. Il s'agit par exemple de l'estimation des paramètres de croissances (des hauteurs, diamètres…) dans le domaine agronomique , dosages des osmoregulateur (proline, glycine betaine)

✓ Les observations intermédiaires : sont également réalisées à un ou à plusieurs moments différents, durant l'expérience. Il s'agit de variables ou de caractéristiques différentes de celle qui doivent être observées au cour de l'expérience (germination, les différentes stades de développement des fruits)

✓ Les observations de contrôle : elles sont relatives aux conditions dans les quelles l'expérience se déroule. Elles doivent portées sur les conditions contrôlés et sur les facteurs constants de l'expérience. On peut citer comme exemple les facteurs contrôles la température, humidité (si l'expérience se déroule dans une chambre de culture).

1. Les paramètres de positions

Ils visent à résumer la zone réels où se trouvent les observations faites sur l'échantillon

1. Moyenne

Soit x1, x2, ..., xp les p valeurs d'une série statistiques discrète, et n1, n2, ..., np les effectifs associés à ces valeurs. Soit N = n1 + n2 + — + np l'effectif total. La moyenne de cette série statistique est alors le nombre

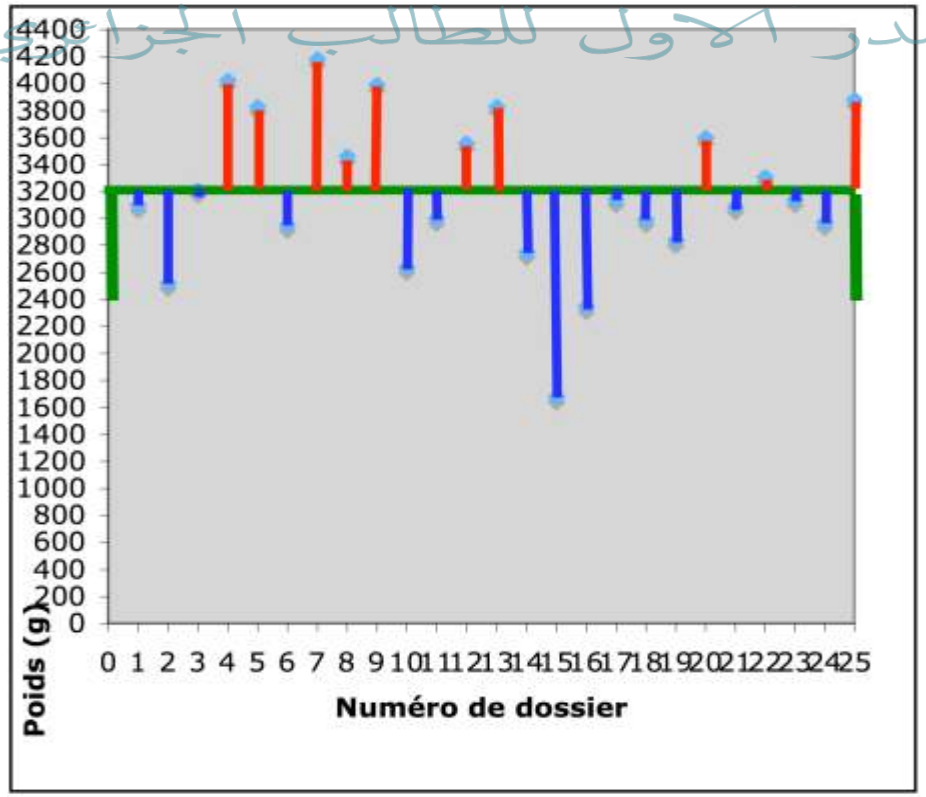
$$\bar{x} = \frac{n_1 x_1 + n_2 x_2 + \ldots + n_p x_p}{N}.$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \bar{x} = \frac{\sum_{j=1}^{k} n_j x_j}{n} = \sum_{j=1}^{k} f_j x_j$$

Avantage : Facile à calculer

Inconvénient : Sensible aux erreurs

Reprenons l'exemple de la variable X « poids de naissance », sur les 25 premiers accouchements $\overline{x}$ = 3201 g



Patiente #	Poids du nouveau-né (en g)
1	3100
2	2520
3	3210
4	4020
5	3830
6	2950
7	4180
8	3460
9	3990
10	2640
11	3000
12	3560
13	3830
14	2750
15	1680
16	2350
17	3140
18	2990
19	2840
20	3600
21	3090
22	3310
23	3140
24	2970
25	3880

La moyenne est la position d'équilibre que prendra une barre rigide forcée à rester horizontale, attachée par des ressorts aux données

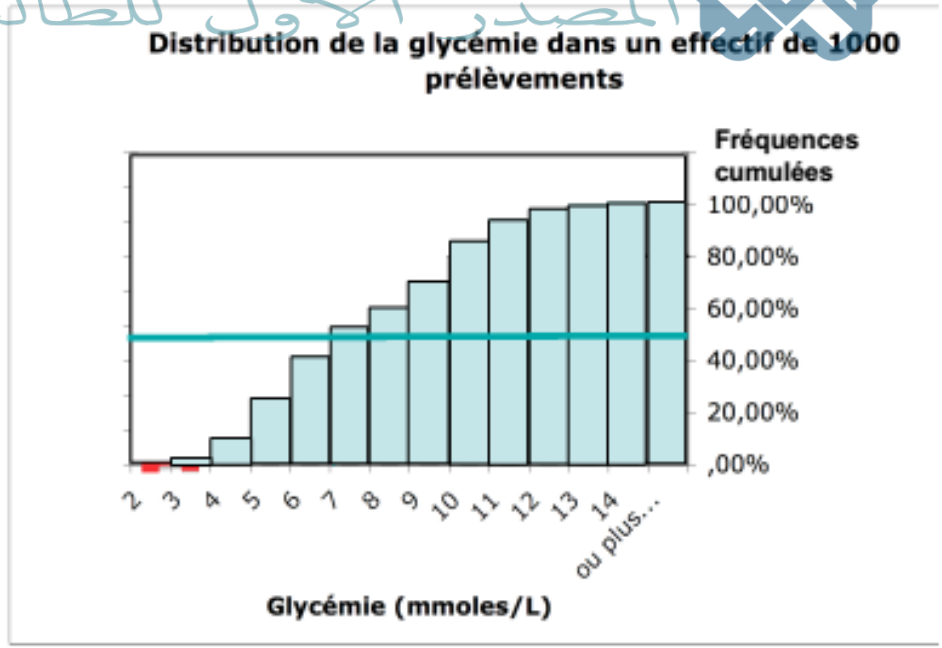
2. MÉDIANE

La médiane d'une série statistique est le nombre qui partage la population en deux parties de même effectif : les individus dont la valeur du caractère est inférieure à la médiane, et les individus dont la valeur du caractère est supérieure à la médiane.

Exemple: Dans une classe de 35 élèves, si les notes sont rangées par ordre croissant, la médiane est la 18eme note (il y en a 17 avant et 17 après). Dans une classe de 34 élèves, si les notes sont rangées par ordre croissant, la médiane est la moyenne de la 17eme note et de la 18eme note (il y en a 17 avant et 17 après).

Avantage de la médiane

- ✓ Peu sensible aux erreurs
- ✓ Facile à comprendre sur la courbe des Fréquences cumulées (exemple : variable X = glycémie, mesurée sur un échantillon de 1000 patients) : médiane = classe [7 ; 8[

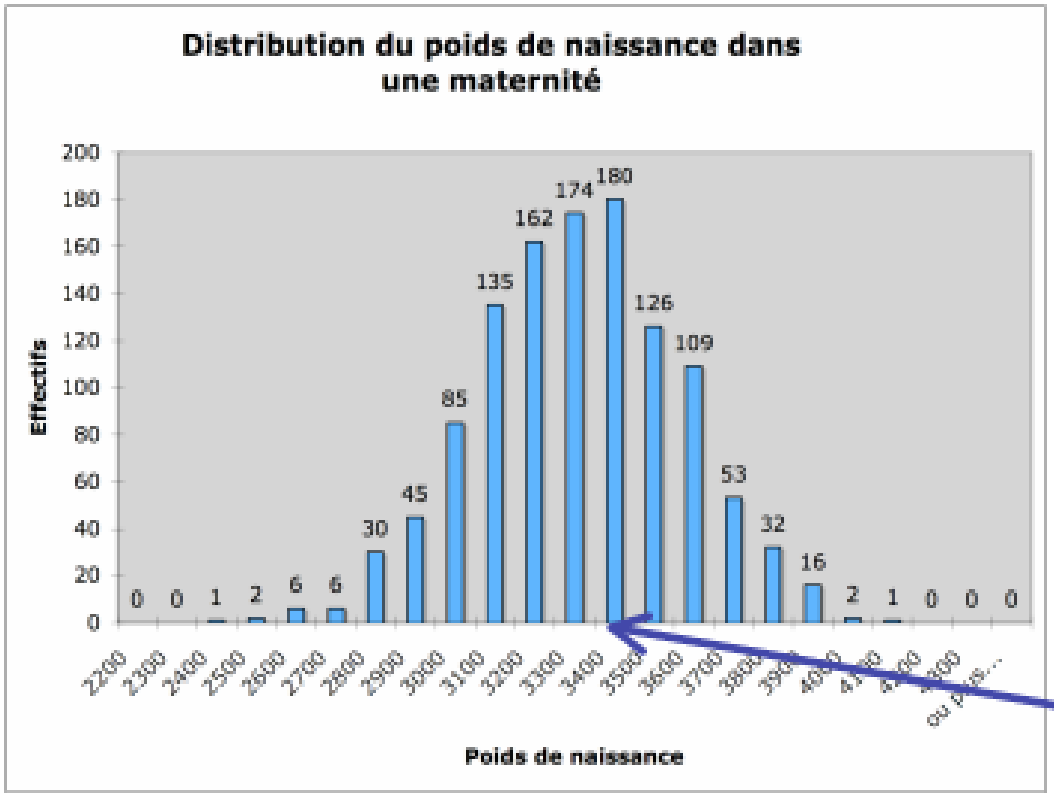


Inconvénient (relatif avec les ordinateurs) :

- ✓ Nécessite de classer les données par ordre

3. MODE OU CLASSE MODALE

Le mode d'une série statistique discrète est la valeur du caractère qui possède le plus gros effectif. La classe modale d'une série statistique continue est la classe qui possède le plus gros effectif.



Mode = maximum local des effectifs = 3400 (ou [3350;3450[)

4. L'ÉTENDUE

L'étendue d'une série statistique quantitative est la différence entre la plus grande et la plus petite des valeurs du caractère.

Remarque : L'étendue est une mesure de la dispersion d'une série statistique, tandis que les moyenne, mode et médiane mesure la position d'une série statistique.

5. LA FRÉQUENCE

La fréquence d'une valeur d'un caractère est le quotient de l'effectif par l'effectif total.

$$f_i = \frac{n_i}{n}$$

Exemple: Considérons la taille des 35 élèves d'une classe de seconde (à compléter) :

Taille (en m)	[1,5; 1,6[	[1,6; 1,7[	[1,7; 1,8[	[1,8; 1,9[	[1,9; 2[
Effectif	5	16		4	1
Fréquence	$\frac{5}{35} \cong 0{,}14$				

2. Les Paramètres de dispersion

Ces paramètres ont pour objectif dans le cas d'un caractère quantitatif de caractériser la variabilité des données dans l'échantillon.

Les indicateurs de dispersion fondamentaux sont la variance observée et l'écart-type observé.

1.La variance observée

Soit un échantillon de n valeurs observées x1, x2, ….,xi,….,xn d'un caractère quantitatif X et soit $\bar{x}$ la moyenne observée. On définit la variance observée notée s2 comme **la moyenne arithmétique des carrés à la moyenne.**

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

Autre formule pour calculer la variance :

$$\mathbf{V} = \frac{1}{N}\left[n_1 x_1^{\,2} + n_2 x_2^{\,2} + \cdots + n_i x_i^{\,2} + \cdots . + n_p x_p^{\,2}\right] - (\bar{x})^2$$

2. L'écart-type

observé correspond à la racine carrée de la variance observée:

$$\sigma_x = \sqrt{V(x)}$$

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

Soit la série statistique définie dans le tableau suivant :

Valeur	x_1	x_2	.....	x_p
Effectif	n_1	n_2	.....	n_p
Fréquences	f_1	f_2		f_p

Effectif total : $N = n_1 + n_2 + \ldots + n_p$ **et** $f_i = \frac{n_i}{N}$

Soit $\overline{x}$ **la moyenne de cette série .**

Le réel V = $\frac{1}{N}[n_1(x_1 - \overline{x})^2 + n_2(x_2 - \overline{x})^2 + \cdots + n_i(x_i - \overline{x})^2 + \cdots + n_p(x_p - \overline{x})^2]$
est appelé variance de cette série statistique.

La racine carrée de la variance $\sigma = \sqrt{V}$ **est l'écart type de cette série.**

La **variance** et **l'écart type** permettent de mesurer la « **dispersion** » des valeurs de la série autour de la moyenne.
Si les valeurs de la série possèdent une unité, l'écart type s'exprime dans la même unité.

Exemples : Calculs de la variance et de l'écart type des séries précédentes

1°) Soit la série statistique répertoriant la taille en mètres de 100 requins blancs

taille (en m)	1,5	2	2,5	3	3,5	4	4,5
Effectif	8	10	25	32	19	4	2

La taille moyenne est :

$$\overline{x} = \frac{1{,}5\times8 + 2\times10 + 2{,}5\times25 + 3\times32 + 3{,}5\times19 + 4\times4 + 4{,}5\times2}{100} = 2{,}82$$

La variance $V = \frac{1{,}5^2\times8 + 2^2\times10 + 2{,}5^2\times25 + 3^2\times32 + 3{,}5^2\times19 + 4^2\times4 + 4{,}5^2\times2}{100} - 2{,}82^2$

$V = 8{,}395 - 7{,}9524 = 0{,}4426$ et $\sigma = \sqrt{0{,}4426} \approx 0{,}665$ m

On veut la moyenne du taux de glucose dans le mélange final de 4 types de mangues :

Concentration ($g.L^{-1}$) X	Moyenne $x_j^\star$	Nb de mangues n_j
$[135, 165[$	150	17
$[165, 180[$	172.5	23
$[180, 195[$	187.5	14
$[195, 225[$	210	8

$$\bar{x} = 174.56 \text{ g.L}^{-1}$$

$$s^2 = \frac{1}{62}\left(17 \times 150^2 + 23 \times 172.5^2 + \ldots\right) - 174.56^2$$

$$= 365.60$$

$$s = \sqrt{365.60} = 19.12 \text{ g.L}^{-1}$$

3. Coefficient de variation

La variance et l'écart-type observée sont des paramètres de dispersion absolue qui mesurent la variation absolue des données indépendamment de l'ordre de grandeur des données.

Le coefficient de variation noté C.V est un indice de dispersion relatif prenant en compte ce biais et est égal à :

$$C.V. = \frac{100s}{\bar{x}}$$

ANALYSE DE LA VARIANCE

ANOVA à un facteur - Introduction

L'analyse de la variance a pour but la comparaison des moyennes de **N** population, à partir d'échantillons aléatoires et indépendants prélevés dans chacune d'elles.

Ces populations sont en général des variantes d'un ou plusieurs facteurs contrôlés de variation (facteurs **A**, **B**, ...).

Conditions d'applications de l'ANOVA

• Les populations étudiées sont normalement distribuées (ou approximativement normale);

• Les variances entre les populations sont identiques

• Les échantillons sont prélevés aléatoirement et indépendamment dans les populations.

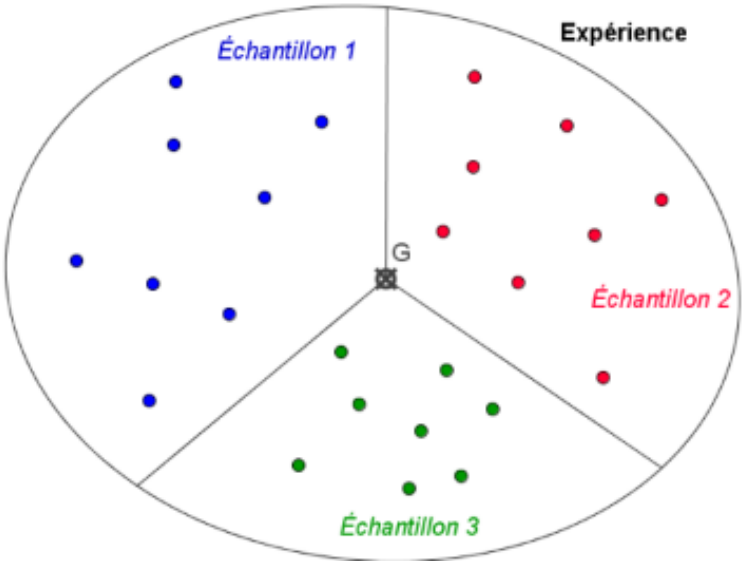
ANOVA à un facteur - **Schématisation** de l'analyse multiple de moyennes

* Expérience avec N échantillons - Données initiales

• Soit une Expérience faisant intervenir **K** échantillons de n_i individus.

•Le nombre total d'individus est $N = \Sigma n_i$

•On calcule la moyenne générale des mesures de l'expérience (**G**).



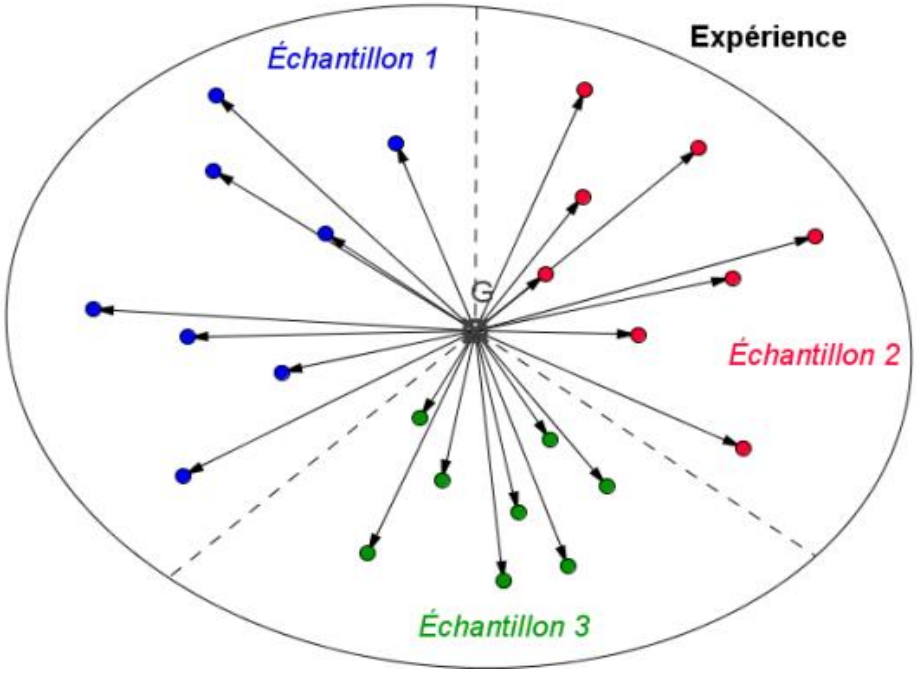
Expériences avec plusieurs échantillons

Variabilité totale

• Variabilité totale au sein de l'expérience (quel que soit l'échantillon) : reflète les écarts de tous les individus par rapport à la moyenne générale (**g**) de l'expérience.

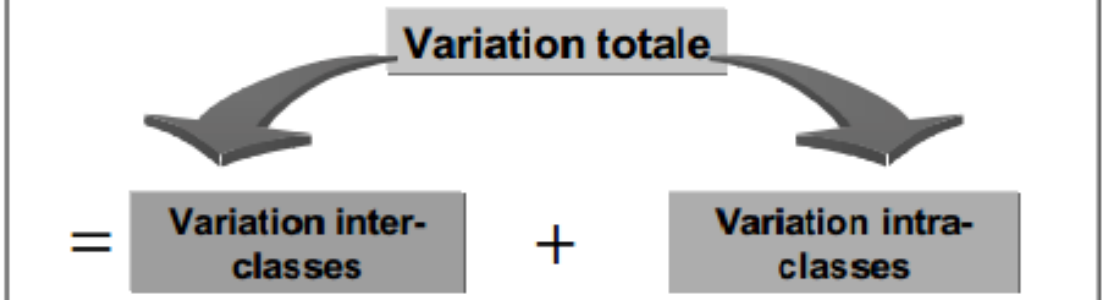
• Calcul de la somme des Carrés des Écarts à la moyenne totale (SCE_T).

• Degrés de liberté (ddl) associés : N-1.



Variabilité totale (toutes les échantillons confondus)

Décomposition de la variation totale



- Somme des carrés du modèle
- Somme des carrés due au facteur
- Sum of Squares Between

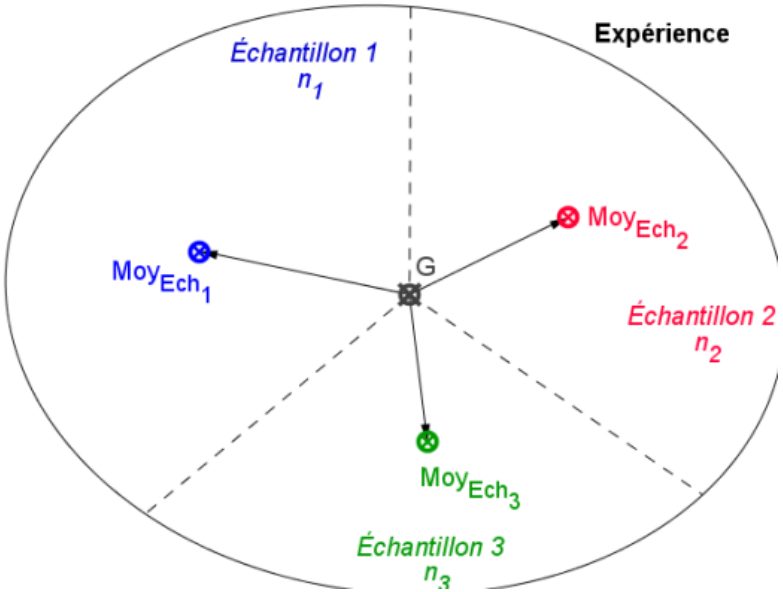
- Somme des carrés des erreurs
- Sum of Squares Within

Variabilité factorielle

Reflète les écarts des moyennes des échantillons (supposées influencées par le facteur étudié) par rapport à la moyenne factorielle.

• Calcul de la Somme des Carrés des Écarts à la moyenne factorielle (SCE_F).

•ddl factorielle associés : k-1.



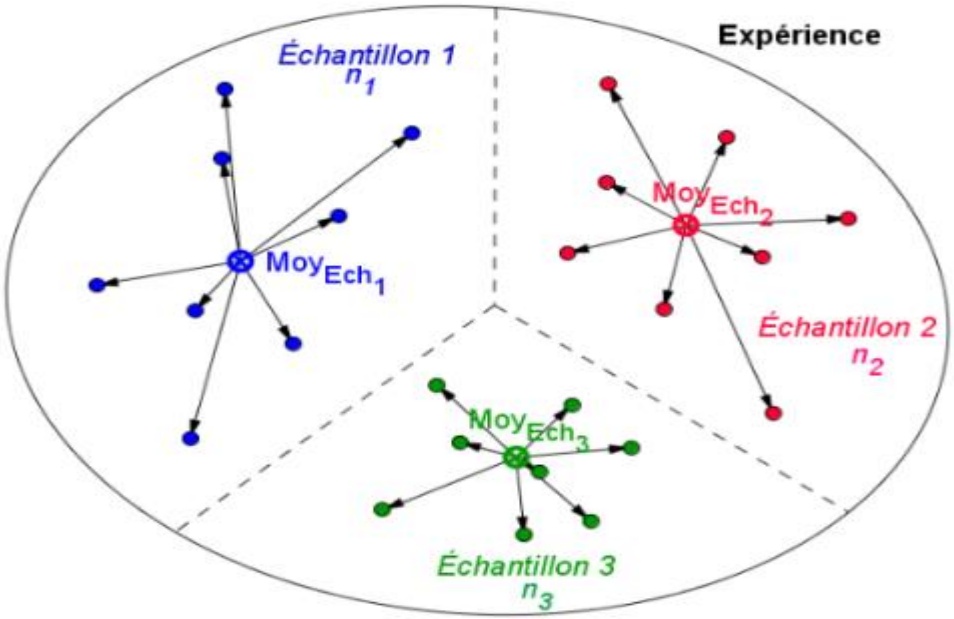
Effet du facteur étudié sur les moyennes des échantillons par rapport à la moyenne générale

Variabilité résiduelle

• Variabilité résiduelle (liée à l'individu) : reflète l'importance des variations individuelles dans chaque échantillon.

• Calcul de la Somme des Carrés des Écarts à la moyenne résiduelle (SCE_R).

• ddl résiduelle associés : N-k.

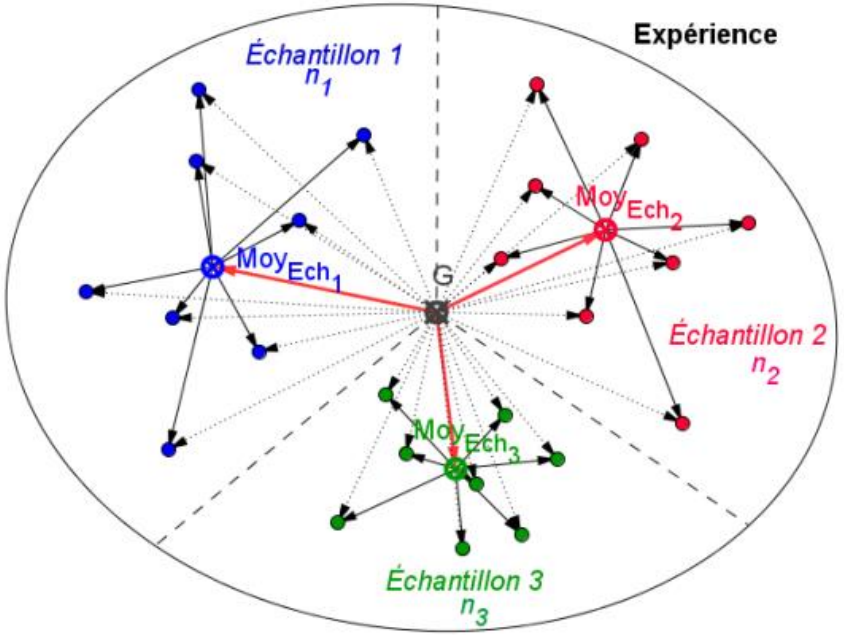


Variabilité intra-groupe (résiduelle)

Bilan

Pour résumer :

- $SCET = SCEF + SCER$
- DDL associés : N-1 = k-1 + N-k.
- On comparera les variabilités factorielle et résiduelle



Représentation combinée de toutes les sources de variabilités

Pour chaque échantillon E_i de taille n_i , on calcule :

- Moyenne
- Moyenne générale
- Variance expérimentale
- Variance factorielle (variance intergroupe) :

  Dispersion des valeurs d'un échantillon à l'autre (influence du facteur)
- Variance résiduelle (variance intragroupe):

  Dispersion des valeurs à l'intérieur des échantillons (variabilité individuelle)

Tableau d'analyse de la variance

Source de variation	Degrés de liberté	Somme Des carrés	Moyenne des carrés (Variance)	F
Facteur	k - 1	SCF	MCF = SCF/(k - 1)	$\frac{MCF}{MCE}$
Erreur	n - k	SCE	MCE = SCE/(n - k)	
Total	n - 1	SCT = SCF+SCE		

Formules (1)

$$SCT = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{x}\right)^2$$

$$SCF = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\bar{x}_i - \bar{x}\right)^2$$

$$SCE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{x}_i\right)^2$$

s