# PROCESSUS STOCHASTIQUE ANOVA

PR. M. DJAMEL MOUSS

Laboratoire d'Automatique et Productique LAP

# ANOVA

# ANalyse Of VAriance

## Position du problème :

Exemple introductif :

- On veut connaître l'effet de trois types de fertilisants sur la croissance des arbres d 'une plantation
- On veut connaître l'effet de la race des vaches laitière sur la production laitière de ces vaches

Les domaines d'études sont variés. L'ANOVA s 'applique dès que :

- on veut monter une expérimentation
- on veut vérifier l 'effet de variables qualitatives sur une variable quantitative

extraire 3 échantillons (groupes) d ’arbres et appliquer chaque **fertilisant** pour chaque échantillon : comparer ensuite les moyennes de **croissance annuelle** des arbres

***Variable d’interet (variable dépendante).***

En cm/an

Autre exemple

- Rendement production laitière
- Taux de virus dans le sang

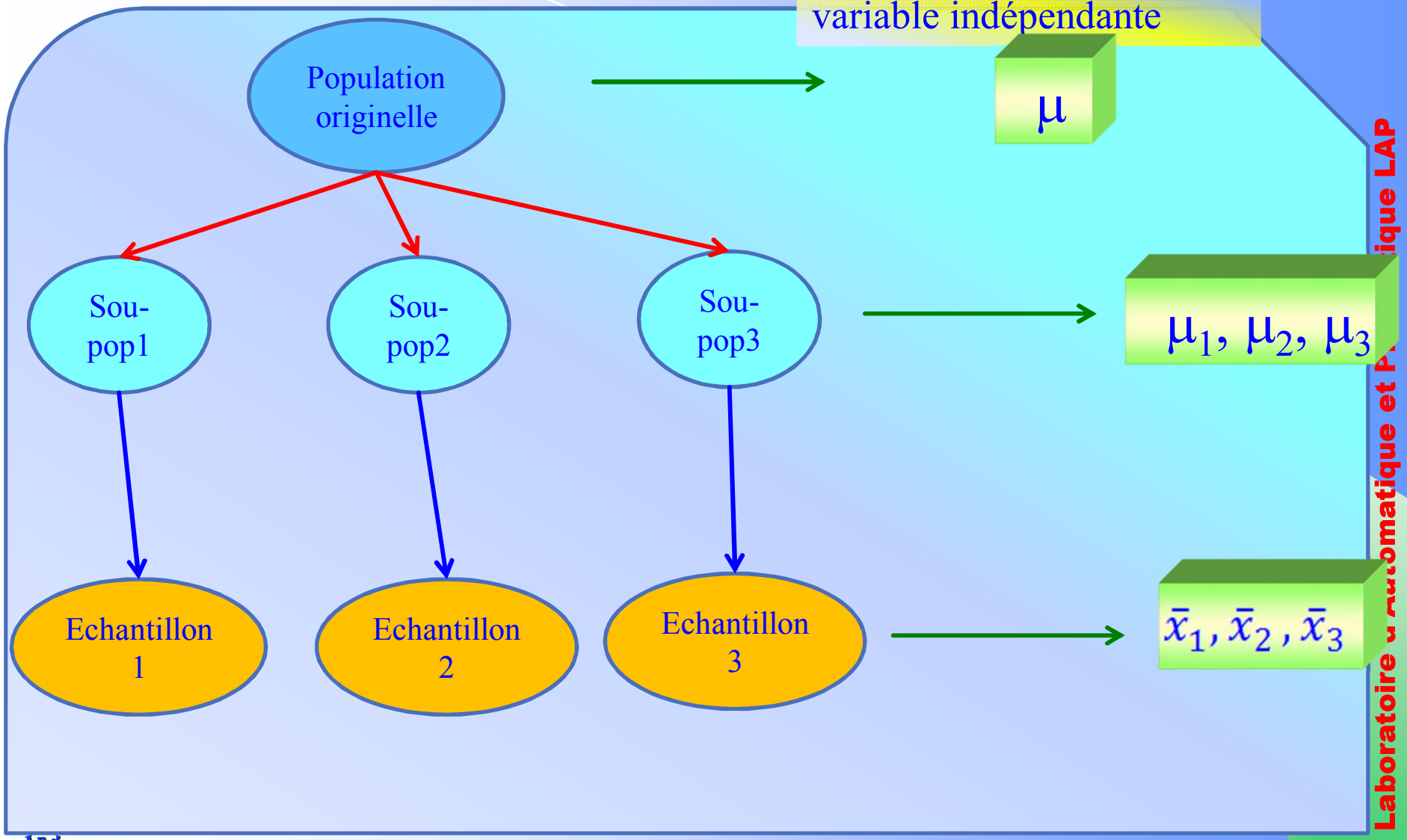
***Facteur (variable indépendante).***

Type de fertilisant

Autre exemple

- Race des vaches
- Traitement médical

# Principe statistique



## *Question ?* Pourquoi ne pas faire de multiples tests *t*?

Par exemple pour 3 groupes (A-B-C), on aura 3 comparaisons (A-B, B-C, A-C), 3 tests *t* : Il y a rapidement trop de comparaisons à faire.

## La solution ANOVA

- Analyser toute la variance.
- Classer la variance en catégories et comparer les catégories.
- Deux catégories principales:
  - la variance entre les groupes (variance intergroupe)
  - la variance à l'intérieur des groupes (variance intra-groupe)

La problématique de l ’ANOVA consiste à utiliser les moyennes observées sur les échantillons pour conclure à des différences significatives sur les moyennes (espérance mathématique) dans les sous-populations

Hypothèses stochastiques

- les échantillons sont issus d’une population normale (gaussienne) : on parle de test paramétrique
- les variances conditionnelles (variances dans chaque sous-population) sont identiques : homoscédasticité
- les sous-échantillons sont indépendants

# Présentation des données

Deux types de tableaux sont disponibles :

## a) Adaptés pour la compréhension du problème et les calculs « à la main »

Etudier la puissance des véhicules selon le type de carburant utilisé

« Puissance »
Variable d'intérêt

| ESSENCE | DIESEL |
|---|---|
| 111 | 64 |
| 111 | 72 |
| 154 | 123 |
| 102 | 123 |
| 115 | 123 |
| 110 | |
| 110 | |
| 110 | |
| 140 | |

Facteur qui prend deux modalités (essence, diesel)

Pour chaque modalité du facteur, on dispose des observation de la variable d'intérêt ( 9 voitures essence, 5 voitures diesel)

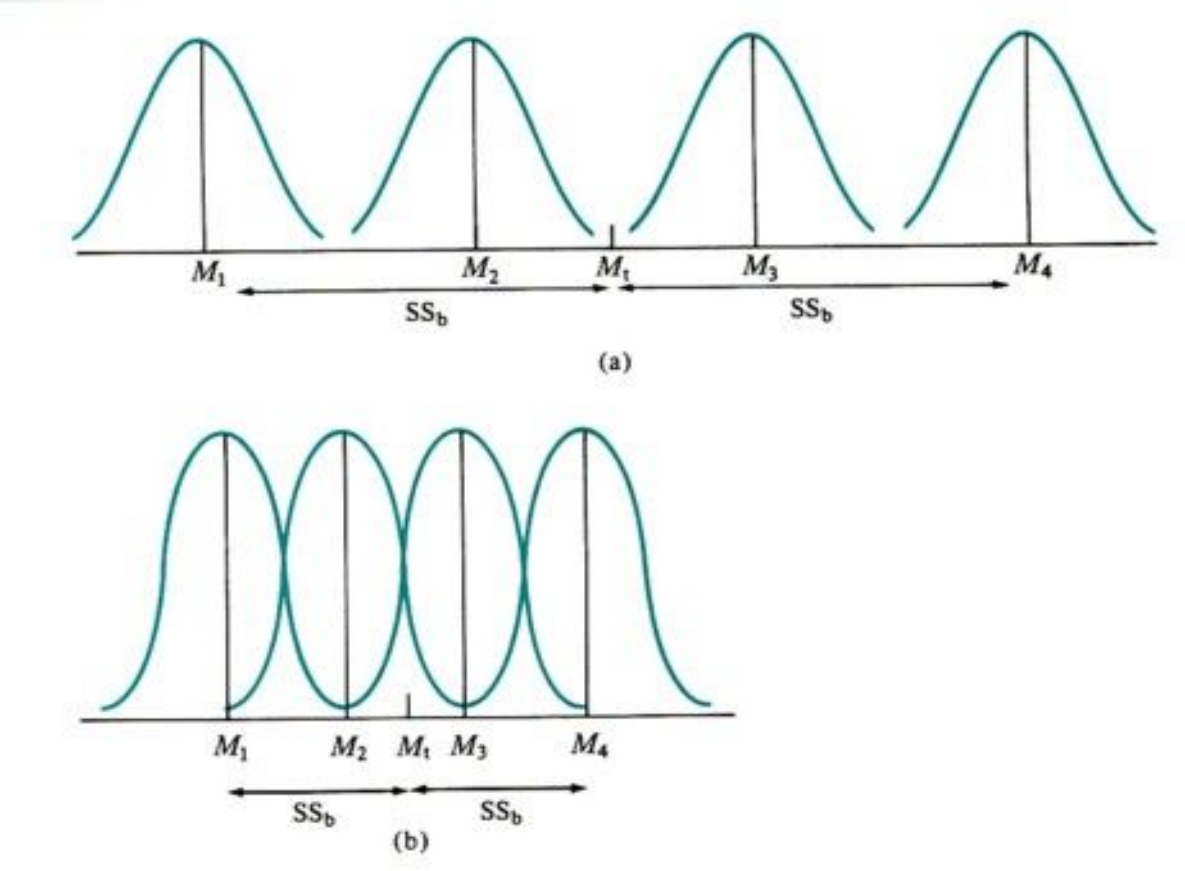
# b) Que l ’on retrouve sur la plupart des logiciels de statistique

On dispose de la liste des observation, à chaque ligne (observation ) on observe la valeur prise par la variable d’intérêt et la valeur prise par le facteur

| Puissance | Carburant |
|---|---|
| 111 | Essence |
| 111 | Essence |
| 154 | Essence |
| 102 | Essence |
| 115 | Essence |
| 110 | Essence |
| 110 | Essence |
| 110 | Essence |
| 140 | Essence |
| 64 | Diesel |
| 72 | Diesel |
| 123 | Diesel |
| 123 | Diesel |
| 123 | Diesel |

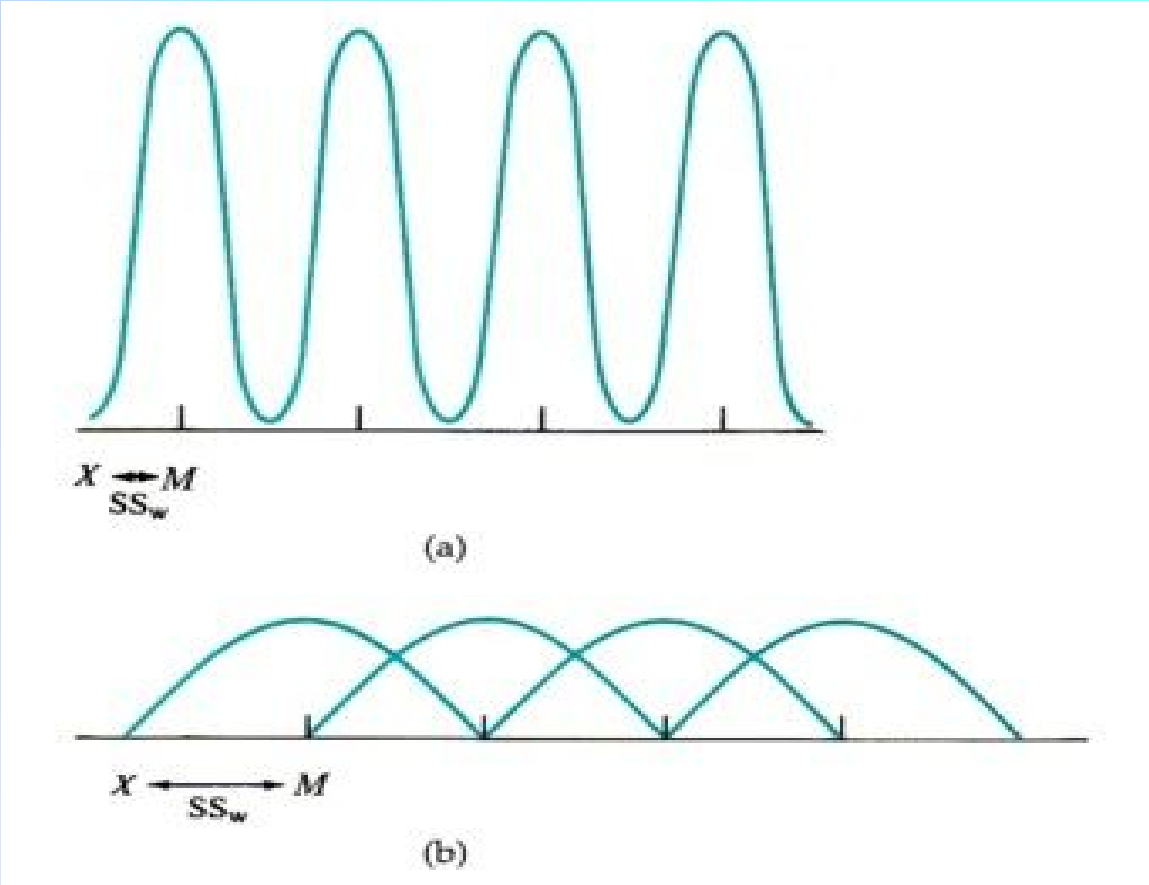
# La variance intergroupe

**Mesure de la variance entre les moyennes de groupes et entre celles-ci et la moyenne totale.**



# La variance intra-groupe

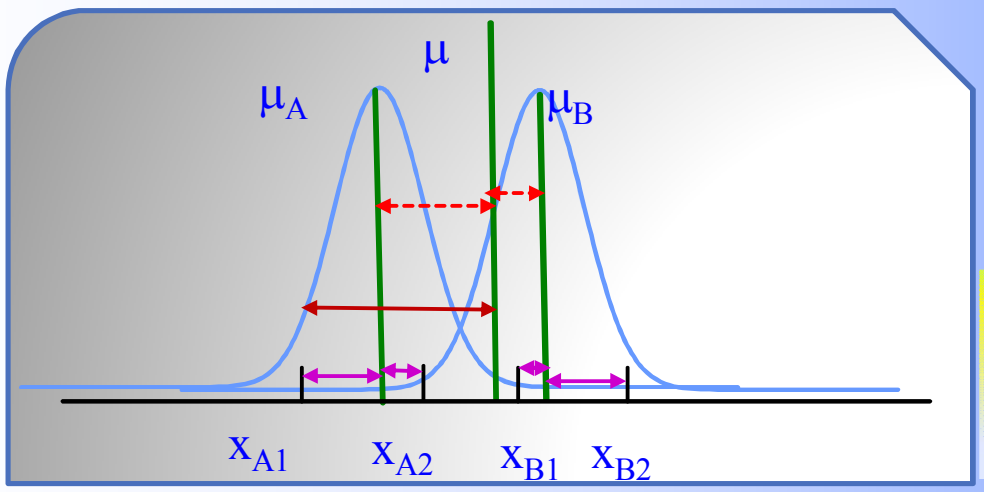
**Mesure de la variance entre les observations et leur moyenne de groupe.**



# Les types de variations

***Variation inter-échantillon :*** Elle est due aux écart entres les moyenne de chaque échantillon et la moyenne générale et qui traduit l'effet du facteur : ***Variation Factorielle***

***Variation intra-échantillon :*** Elle cumule les écart de chaque valeur individuelle da la variable à leur moyenne d'échantillon . Cette dispersion provient des fluctuation aléatoire d'échantillonnage: ***Variation Résiduelle***

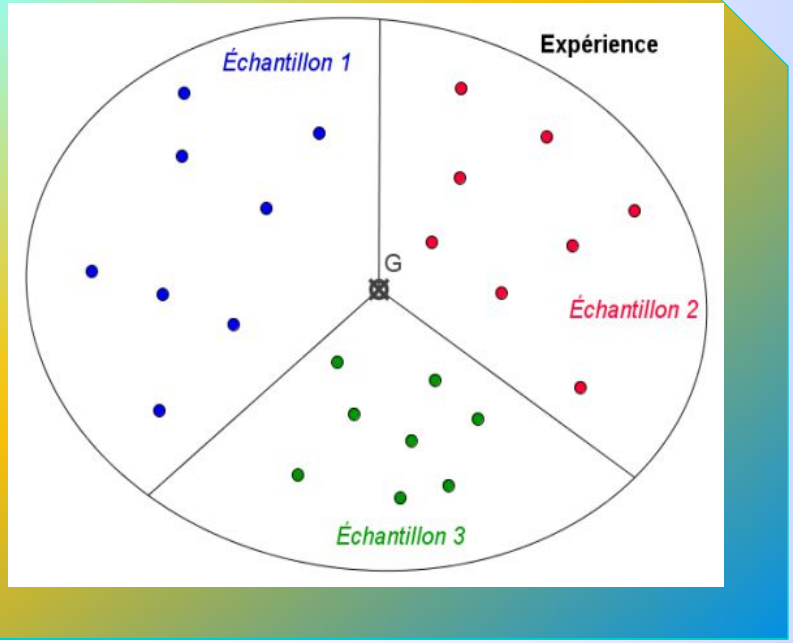


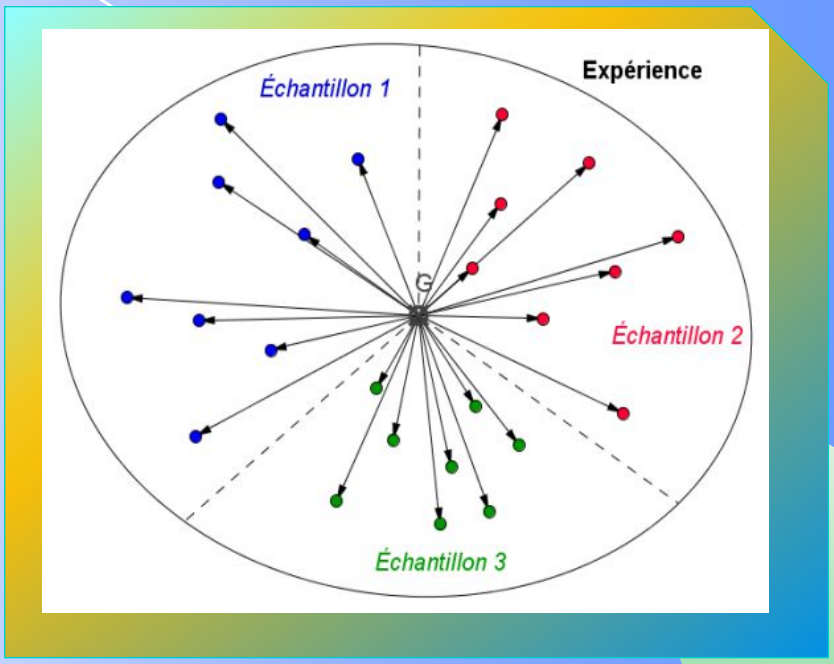

***Variation intra***

***Variation inter***

***Variation total***

# Schématiquement nous aurons sur un exemple à 3 échantillon




***Expérience*** **:** p (dans ce cas 3) échantillons

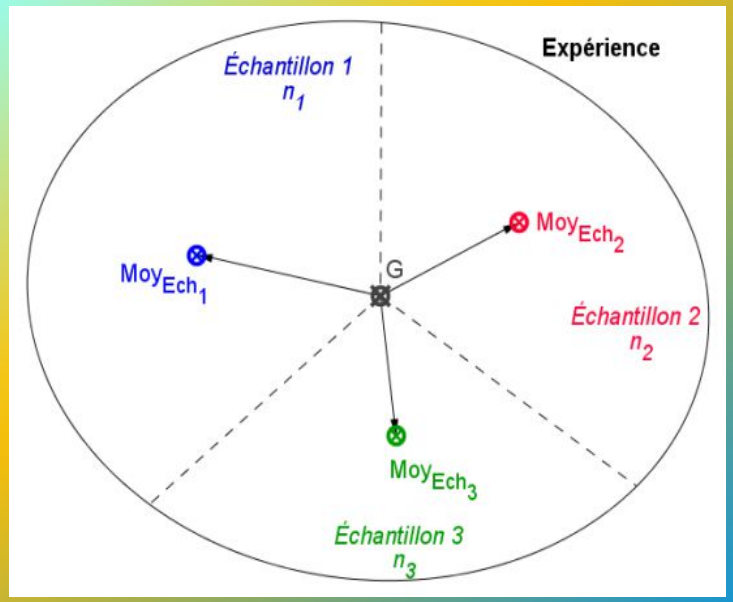
n: taille de la population d'étude.

***Variabilité totale au sein de l'expérience*** :
Elle reflète les écarts de tous les individus par rapport à la moyenne générale de l'expérience.

Degrés de liberté (DDL) associés : n-1.

$SCE_T$

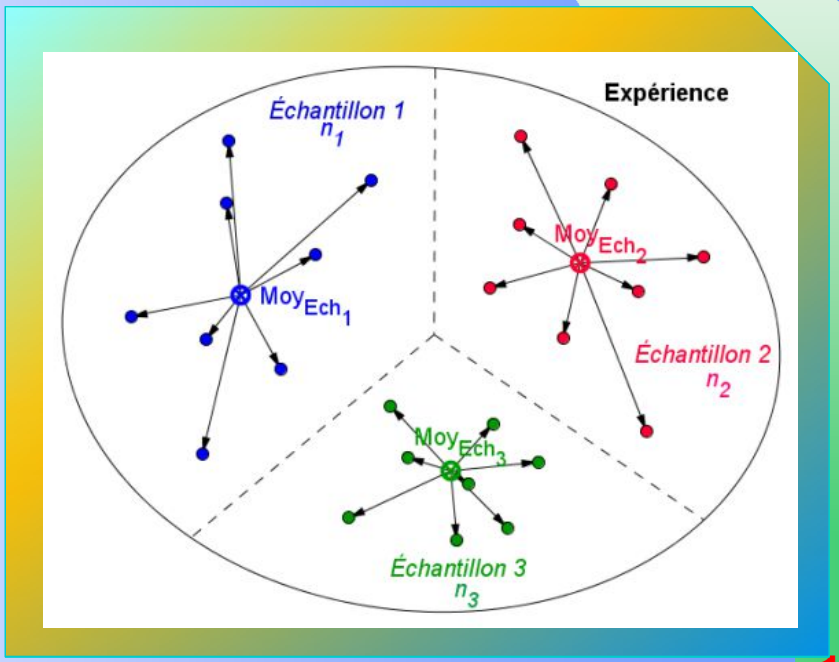
# Schématiquement nous aurons sur un exemple à 3 échantillon




***Variabilité factorielle :*** Elle reflète les écarts des moyennes des échantillons par rapport à la moyenne générale de l'expérience.

Degrés de liberté (DDL) associés : p-1.

$SCE_F$




***Variabilité résiduelle:*** Elle reflète l'importance des variations individuelles dans chaque échantillon.

Degrés de liberté (DDL) associés : n-p.

$SCE_R$

# Position du Problème et notation

Population :P ⟶ Subdivision en sous population $P_1, P_2, \ldots\ldots P_p$,

Facteur à étudier :A ⟶ Ayant n modalité $A_1, A_2, \ldots\ldots A_p$,

Variable d'intérêt :X ⟶ Ayant une moyenne $\mu$ ce qui donne dans chaque sous population $\mu_1, \mu_2, \ldots\ldots \mu_p$,

Echantillon :E⟶ Subdivision en sous échantillon $E_1, E_2, \ldots\ldots E_p$, et de taille $n_1, n_2, \ldots\ldots n_p$

$$n = \sum_{j=1}^{p} n_j$$

On détermine pour ces échantillons les moyennes empiriques

$$\bar{x} \quad et \quad \bar{x}_1, \bar{x}_2, \ldots \ldots, \bar{x}_p$$

L'ANOVA consiste à construire le test d'hypothèse

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_p = \mu$$
$$H_1 : \exists j, \ \mu_j \neq \mu$$

Le facteur A n'a aucune influence sur la variable dépendante

Sur l'échantillon nous pouvons calculer alors :

Moyenne Conditionnelle
( pour chaque facteur)

$$\bar{x}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij}$$

Moyenne Globale (tous
Facteur confondus)

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{p} n_j \bar{x}_j$$

Parallèlement à la présentation théorique traitons un exemple concret :

Présentation : Nous allons comparer 3 machines dont l'action est de remplir des flacons . On prélève sur chaque machine un échantillon aléatoire de la production obtenue au cours de 5 périodes différentes ( On pose la Variable Aléatoire : X = volume du flacon).Les données sont représenté sur le tableau suivant

| Machine 1 | Machine 2 | Machine 3 |
|---|---|---|
| 47 | 55 | 54 |
| 53 | 54 | 50 |
| 49 | 58 | 51 |
| 50 | 61 | 51 |
| 46 | 52 | 49 |
| 245 | 280 | 255 |

## Moyenne Conditionnelle( pour chaque facteur)

$$\bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{i1} = \frac{1}{5}(245) = 49$$

$$\bar{x}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{i2} = \frac{1}{5}(280) = 56$$

$$\bar{x}_3 = \frac{1}{n_3}\sum_{i=1}^{n_3} x_{i3} = \frac{1}{5}(255) = 51$$

## Moyenne Globale (tous Facteur confondus)

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{p} n_j \bar{x}_j = \frac{1}{15}(5 * 49 + 5 * 56 + 5 * 51) = \frac{780}{15} = 52$$

Variation Totale

**Variation Interclasses** + **Variation Intra-classes**

$$x_{ij} - \bar{x} = (\bar{x}_j - \bar{x}) + (x_{ij} - \bar{x}_j)$$

*Écart à la moyenne global*

*Écart entre les groupes (définis par les facteurs)*

*Écart à l'intérieur des groupes*

$$\sum_{j=1}^{p}\sum_{i=1}^{n_j}(x_{ij} - \bar{x})^2 = \sum_{j=1}^{p}(\bar{x}_j - \bar{x})^2 + \sum_{j=1}^{p}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2$$

SCT: Somme des carré totaux
Exprime la variabilité totale des observation

SCE: Somme des carré expliqués
Exprime la variabilité expliquée soit la variation que le facteur explique

SCR: Somme des carré résiduels
Exprime la variabilité résiduelle soit la variation que le facteur n'arrive pas à expliquer

## $SCE_R$

$$\sum_{j=1}^{p}\sum_{i=1}^{n_j}\left(x_{ij}-\bar{x}_j\right)^2 = (47-49)^2+(53-49)^2+(49-49)^2+(50-49)^2$$

$$= 30+50+14 = 94$$

## $SCE_F$

$$\sum_{j=1}^{p}\left(\bar{x}_j-\bar{x}\right)^2 = 5*(49-52)^2+5*(56-52)^2+5*(51-52)^2$$

$$= 45+80+5 = 130$$

# TABLEAU DE L'ANOVA

| Variation | SCE | ddl | CM | F |
|---|---|---|---|---|
| Factorielle | $SCE_F = (p-1) * S_F^2$ | $(p-1)$ | $CM_F = S_F^2$ | $F = \frac{S_F^2}{S_R^2}$ |
| Résiduelle | $SCE_R = (n-p) * S_R^2$ | $(n-p)$ | $CM_R = S_R^2$ | |
| Totale | $SCE_T = (n-1) * S_T^2$ | $(n-1)$ | $CM_T = S_T^2$ | |

## Pour notre exemple

| Variation | SCE | ddl | CM | F |
|---|---|---|---|---|
| Factorielle | 130 | 2 | 75 | $F = \frac{75}{7.83}$ |
| Résiduelle | 94 | 12 | 47 | |
| Totale | 224 | 14 | 56 | $= 9.58$ |

# Prise de décision

La variable F représentant le rapport : $F = \frac{S_F^2}{S_R^2}$ suit une distribution de de Fisher de ddl:

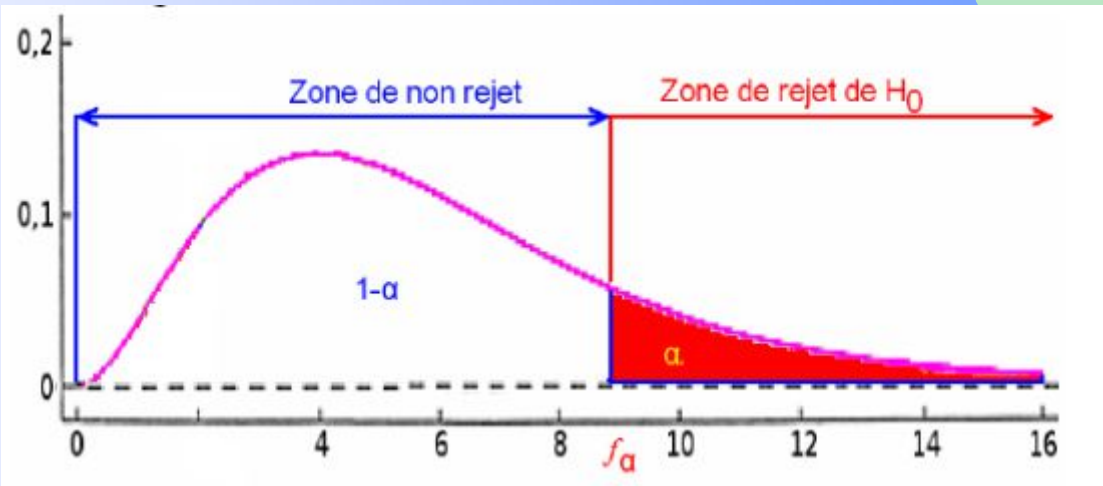
$\nu_1$ : ddl du numérateur soit p-1 .

$\nu_2$ : ddl du dénominateur soit n-p .

$$F \rightsquigarrow \Im_{\alpha,p-1,n-p}$$

On rejette l'hypothèse $H_0$ si F est supérieur à la valeur donné par le seuil de signification choisit :

$H_0$ rejeté si $F > \Im_{\alpha,p-1,n-p}$



Pour notre exemple

$\nu_1$ : ddl du numérateur soit p-1 = 3-1 2

$\nu_2$ : ddl du dénominateur soit n-p = 5-3=2

$\Im_{0.05,2,2} = 19.00$

Décision : Comme F = 9.58 < 19.00 . On adopte alors $H_0$

# Exemple

Pour étudier l'influence du facteur « intensité du bruit environnant » sur la capacité d'un sujet à résoudre un problème. On prend 24 sujet répartis en 4 groupes. Ces 24 sujets passent un contrôle, chaque groupe dans uns salle avec un niveau sonore différent diffusé dans chaque salle. La note à l'épreuve constitue la variable réponse. Nous obtenons les données suivante

-Niveau sonore désigné par 1-2-3et 4
-Les groupes sont constitués par 4, 8 6 et 6 individus

| | Niveau sonore | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| Note | 62 | 56 | 63 | 68 |
| | 60 | 62 | 67 | 66 |
| | 63 | 60 | 71 | 71 |
| | 59 | 61 | 64 | 67 |
| | | 63 | 65 | 68 |
| | | 64 | 66 | 68 |
| | | 63 | | |
| | | 59 | | |

# Le tableau des calculs donne:

| Note | Niveau sonore | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | |
| | 62 | 56 | 63 | 68 | |
| | 60 | 62 | 67 | 66 | |
| | 63 | 60 | 71 | 71 | |
| | 59 | 61 | 64 | 67 | |
| | | 63 | 65 | 68 | |
| | | 64 | 66 | 68 | |
| | | 63 | | | |
| | | 59 | | | |
| $n_j$ | 4 | 8 | 6 | 6 | |
| $\sum_{i=1}^{n_j} x_{ij}$ | 244 | 488 | 396 | 408 | |
| $\overline{x}_j$ | 61 | 61 | 66 | 68 | |
| $\overline{x}$ | 64 | | | | |
| $\sum_{i=1}^{n_j} (x_{ij} - \overline{x}_j)^2$ | 10 | 48 | 40 | 14 | 112 |
| $n_j (\overline{x}_j - \overline{x})^2$ | 36 | 72 | 24 | 96 | 228 |

| Variation | SCE | ddl | CM | F |
|---|---|---|---|---|
| Factorielle | 228 | 3 | 37.3 | $F = \frac{37.3}{11.4} = 3.27$ |
| Résiduelle | 112 | 20 | 11.4 | |
| Totale | 340 | 23 | 14.78 | |

Dans ce cas

$\nu_1$ : ddl du numérateur soit p-1 = 4-1 = 3

$\nu_2$ : ddl du dénominateur soit n-p = 24-4=20

$\Im_{0.05,3,20} = 8.67$

Décision : Comme F = 3.27 < 8.67 . On adopte alors $H_0$