

**ANOVA à un facteur**  
**The analysis of variance one way**

**Introduction:**

Dans le cadre des tests d'hypothèses , nous avons émis des hypothèses concernant la moyenne d'une population puis comparé les moyennes de deux populations.

Maintenant, on va s'intéresser à la comparaison de plusieurs moyennes à partir d'échantillons aléatoires et indépendants prélevés dans chacune des populations avec des variances égales.

Pour cela nous avons recours à la méthode de l'analyse de la variance (ANOVA)

**Position du problème :**

On veut connaitre l'effet de trois types de fertilisants sur la croissance des arbres d'une plantation

1/principe de l'expérimentation :

- Extraire 3 échantillons (groupes ) d'arbres et appliquer chaque fertilisant pour chaque échantillon : comparer ensuite les moyennes de croissance annuelle des arbres

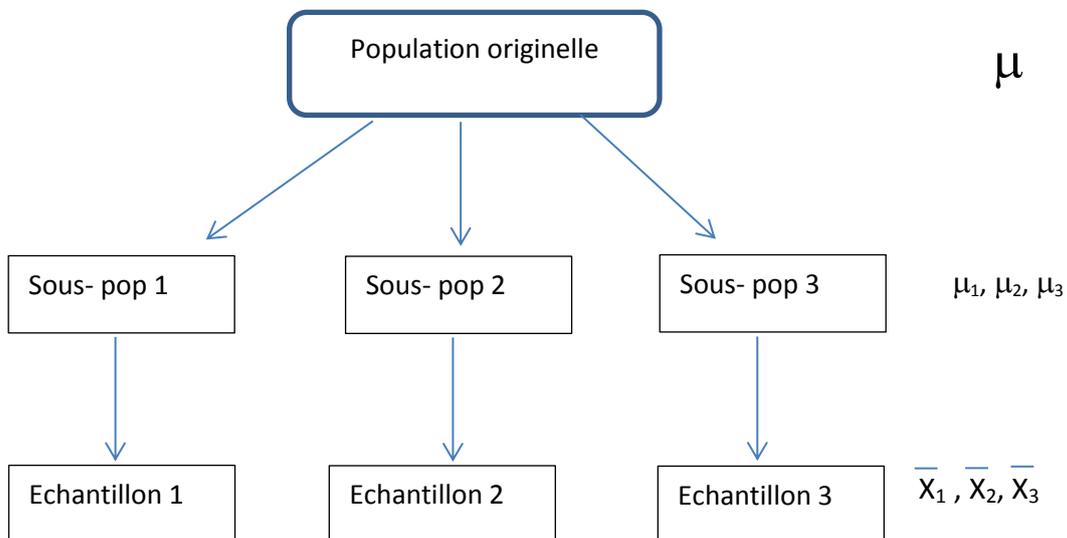
Variable d'intérêt(variable dépendante)  
Exemple : croissance annuelle

facteur (variable indépendante)  
Exemple : type de fertilisant

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

2/ principe statistique :



La problématique de l'ANOVA consiste à utiliser les moyennes observées sur les échantillons pour conclure à des différences significatives sur les moyennes dans les sous-populations.

### Objectif de l'ANOVA :

est de tester l'influence d'un facteur ou plusieurs sur une variable dépendante en d'autres termes cette méthode nous mène à voir l'indépendance d'une variable quantitative avec une ou deux ou plus variables qualitatives .

On a alors ANOVA à un facteur  $\longrightarrow$  analyse bivariée

ou ANOVA à deux facteurs et plus  $\longrightarrow$  analyse multivariée

Le but de cette méthode est la comparaison des moyennes k populations à partir d'échantillons aléatoires et indépendants prélevés dans chacune d'elles

Pour ce chapitre , on s'intéresse à l' ANOVA à un facteur

### Exemple illustratif :

Ech1	Ech2	Ech3
3	5	5
2	3	6
1	4	7

Les étapes de calcul de la procédure :

1/ on calcule la moyenne pour chaque échantillon (série)

$$\bar{X}_1 = (3+2+1)/3 = 2 \quad ; \quad \bar{X}_2 = 4 \quad ; \quad \bar{X}_3 = 6$$

2/ on calcule la moyenne générale de l'ensemble des séries :

$$\bar{X} = (2+4+6) / 3 = 4$$

3/ on calcule la variation totale sous forme des carrés des écarts à la moyenne générale :

$$SCET = (3-4)^2 + (2-4)^2 + (1-4)^2 + (5-4)^2 + (3-4)^2 + (4-4)^2 + (5-4)^2 + (6-4)^2 + (7-4)^2$$

$$SCT = 30 \longrightarrow D.D.L = n-1 = 9-1 = 8$$

4/ on calcule la variation des carrés des écarts à l'intérieur des classes ( ou groupes ou modalités) ou variation intra colonne ou résiduelle :

$$SCEintra = (3-2)^2 + (2-2)^2 + ((1-2)^2 + (5-4)^2 + (3-4)^2 + (4-4)^2 + (5-6)^2 + (6-6)^2 + (7-6)^2$$

$$SCEintra = 6 \longrightarrow D.D.L = n - k = 9-3 = 6$$

5/ on calcule la variation des carrés des écarts entre les classes ou inter colonne ou factorielle :

$$SCE_{inter} = (2-4)^2 + (2-4)^2 + (2-4)^2 + (4-4)^2 + (4-4)^2 + (4-4)^2 + (6-4)^2 + (6-4)^2 + (6-4)^2$$

$$SCE_{inter} = 24 \longrightarrow D.D.L = k-1 = 3-1 = 2$$

On a alors :  $SCET = SCE_{intra} + SCE_{inter}$

$$30 = 6 + 24$$

$$D.D.L \quad 8 = 6 + 2$$

On utilise ces résultats d'un exemple qui teste l'impact de 3 aliments A1 ; A2 ; A3 sur l'étude d'individus pour cela on va élaborer un test statistique de comparaison de 3 moyennes

Question :  $H_0$  : l'aliment n'a pas d'impact  $\mu_1 = \mu_2 = \mu_3$

$H_1$  : l'aliment n'a pas d'impact

$$\text{La statistique du test : } F = \frac{SCE_{inter}/(k-1)}{SCE_{intra}/(n-k)} = \frac{24/2}{6/6} = 12$$

Pour un risque  $\alpha = 10\%$  sur la table de Fischer, on lit  $F(\alpha, k-1, n-k) = F(0.01 ; 2 ; 6) = 3.46$

Or  $F > F_\alpha$  donc on rejette  $H_0$

Conclusion : l'alimentation a un impact sur l'étude

### Conditions d'application de l'ANOVA :

1/ Indépendance : les k échantillons comparés sont indépendants

2/ Normalité : la variable étudiée X suit une loi normale dans les k populations

3/ Homoscédasticité : les k populations comparées ont la même variance en d'autres termes le facteur F agit seulement sur la moyenne de la variable X et ne change pas sa variance

Si les variances sont homogènes on peut comparer les moyennes sinon on ne peut comparer des échantillons qui ne varient pas de la même manière

### Pratique de l'analyse de variance :

- Soit une expérience faisant intervenir k échantillons de  $n_i$  individus
- Le nombre total d'individus est :  $n = \sum_{i=1}^k n_i$

Les hypothèses du test statistique :

$H_0$  :  $\mu_1 = \mu_2 = \dots = \mu_k$  ou il y n'a pas d'impact du facteur F sur la variable quantitative X

$H_1$  : il existe au moins  $i, j$  ;  $\mu_i \neq \mu_j$  ou il y a impact du facteur

Les étapes à suivre pour ce test :

1/ pour chaque échantillon  $E_i$  de taille  $n_i$ , on calcule :

$$\text{Moyenne : } \bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

$$\text{Variance : } S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

2/ pour l'ensemble de l'expérience , on calcule :

$$\text{Moyenne générale : } \bar{x} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{ni} x_{ij} = \frac{1}{n} \sum_{i=1}^k ni \bar{x}_i$$

$$\text{Variance totale : } s^2 = \frac{1}{n-1} \sum_{i=1}^k \sum_{j=1}^{ni} (x_{ij} - \bar{x})^2$$

3/ Calcul de la variance résiduelle (ou variance intra –groupe):

La dispersion des valeurs à l'intérieur des échantillons

$$S^2_R = S^2_{\text{intra}} = \frac{1}{n-k} \sum_{i=1}^k (ni - 1) S^2_i = \frac{1}{n-k} \text{SCE}_R$$

4/ calcul de la variance factorielle (ou variance inter-groupe) :

Dispersion des valeurs d'un échantillon à l'autre (influence du facteur)

$$S^2_F = S^2_{\text{inter}} = \frac{1}{k-1} \sum_{i=1}^k ni (\bar{x}_i - \bar{x})^2 = \frac{1}{k-1} \text{SCE}_F$$

**Théorème de l'analyse de la variance:**

variation	Somme des carrés des écarts (SCE)	Degré de liberté(d.d.l)	La statistique du test
Factorielle	$\text{SCE}_F = (k-1) S^2_F$	K - 1	$F = \frac{S^2_F}{S^2_R}$
Résiduelle	$\text{SCE}_R = (n-k) S^2_R$	n- k	
totale	$\text{SCE}_T = (n-1) S^2_T$	n- 1	

$$\text{SCE}_T = \text{SCE}_F + \text{SCE}_R$$

Sous l'hypothèse  $H_0$  , la variable F suit une loi de Fischer –Snedecor avec  $\eta_1 = k-1$  et  $\eta_2 = n-k$  pour d.d.l

Décision :

1/ Si  $F > F_\alpha$  alors on rejette  $H_0$   $\longrightarrow$  on attribue une influence significative au facteur étudié  
et les moyennes diffèrent significativement

2/ Si  $F < F_\alpha$  alors on accepte  $H_0$   $\longrightarrow$  il n' a pas d'influence du facteur étudié  
et les moyennes ne diffèrent pas significativement

Remarque :

On ne peut pas remplacer une ANOVA par une série de tests de Student (test t) il y aura une inflation de l'erreur

Quand on rejette  $H_0$ , pour savoir quelles sont les moyennes significativement différentes plusieurs méthodes associées à l'ANOVA, sont proposées : méthode de Bonferroni, méthode de Tuckey, méthode de Scheffe ; cette dernière est la plus utilisée

### **Méthode de Scheffe**

Elle repose sur le test des contrastes

#### **Définition :**

On appelle contraste noté C une somme pondérée de moyennes

$$C = c_1\bar{x}_1 + c_2\bar{x}_2 + \dots + c_k\bar{x}_k$$

$$\text{Avec } \sum_{i=1}^k ci = 0 \text{ et } \sum_{i=1}^k |ci| = 2$$

Un contraste permet de comparer une moyenne avec une autre moyenne, un ensemble de moyennes à un autre ensemble de moyennes

Exemple sur le contraste :

On suppose que le test d'ANOVA nous conduit à conclure que 4 moyennes sont différentes dans l'ensemble

→ pour comparer  $m_1$  avec  $m_2$  on pose :  $c_1=1$  ;  $c_2=-1$  ;  $c_3=0$  et  $c_4=0$  puis on teste si le contraste  $C=m_1-m_2$  est différent de 0

→ pour comparer  $m_1$  et  $m_2$  avec  $m_3$  et  $m_4$  et on pose :  $c_1=1/2$  ;  $c_2=1/2$  ;  $c_3=-1/2$  ;  $c_4=-1/2$  et on teste  $C = ((m_1+m_2)/2) - ((m_3+m_4)/2)$

#### **Test de Scheffe :**

$$\rightarrow \text{Si } |C| > \sqrt{(k-1)F_{\alpha} S^2_r \sum_{i=1}^k \frac{ci^2}{ni}} \text{ alors le contraste est significativement différent de 0}$$

Avec k : le nombre d'échantillons

$F_{\alpha}$  : la valeur tabulée de Fisher avec k-1 (colonne) et n-k (ligne) d.d.l pour un risque  $\alpha$

$n_i$  : la taille de l'échantillon  $E_i$

$S^2_r$  : la variance résiduelle ou intra-colonne

→ Sinon le contraste n'est pas significativement différent de 0

#### **Exemple:**

On veut comparer trois traitements A, B, C chez des sujets atteints d'une certaine forme de leucémie. Pour cela, on constitue un tirage au sort de 3 groupes de leucémiques qu'on traite chacun par l'un des 3 traitements. On obtient les résultats suivants (avec X représente la durée de rémission exprimée en mois)

traitement	A	B	C
Taille ni	86	242	142
$\sum x$	1161	2904	1750
$\sum x^2$	16100	36500	22340

Comparer les 3 traitements

Solution : il s'agit de comparer les durées moyennes de rémission des 3 traitements

On suppose que la condition de normalité des durées de rémission est vérifiée et la condition d'égalité des variances est satisfaite

- Moyennes et variances pour chaque traitement :

traitement A :  $\bar{x}_1 = 1161/86 = 13.5$      $S^2_1 = 1/86 (\sum x^2_{11}) - \bar{x}^2_1 = (16100/86) - 13.5^2 = 4.95$

traitement B :  $\bar{x}_2 = 2904/242 = 12$      $S^2_2 = 1/242 (\sum x^2_{12}) - \bar{x}^2_2 = (36500/242) - 12^2 = 6.82$

traitement C :  $\bar{x}_3 = 1750/142 = 12.32$      $S^2_3 = 1/142 (\sum x^2_{13}) - \bar{x}^2_3 = (22340/142) - 12.32^2 = 5.54$

- Variance inter-colonne (factorielle):

$$S^2_F = (1/k-1) \sum_{i=1}^k ni (x_i - \bar{x})^2 = 71.65$$

- Variance intra-colonne (résiduelle):

$$S^2_R = (1/n-k) \sum_{i=1}^k nisi^2 = 6.13$$

- Tableau d'analyse de la variance à un facteur

Source de variation	SCE	D.D.L	Carrés moyens
Intergroupe(factorielle)	$SCE_F = 143.30$	$k-1=3-1=2$	$S^2_F = 71.65$
Intragroupe(résiduelle)	$SCE_R = 2862.82$	$n-k= 470-3=467$	$S^2_R = 6.13$
totale	$SCE_T = 3006.2$	$n=470$	

- Test de Fischer :

$H_0$  : les 3 traitements ont la même efficacité

$H_1$  : les 3 traitements n'ont pas la même efficacité

$$F = S^2_F / S^2_R = 71.65 / 6.13 = 11.69$$

La table de Fischer donne :  $F_{2;467;5\%} = 3 = F_\alpha$

Comme  $F > F_\alpha$  alors on rejette  $H_0$

Du fait que les moyennes diffèrent significativement dans l'ensemble, on va comparer les moyennes 2 à 2 par le test de Scheffe

- On range les traitements par ordre croissant

	B	C	A
Moy	12	12.32	13.5

On veut comparer la moyenne 2 avec la moyenne 3 et la moyenne 3 avec la moyenne 1 on construit les contrastes :

$$B \text{ vs } C : c_1 = \bar{x}_2 - \bar{x}_3 = -0.32 ; (c_1=0 ; c_2=1 ; c_3=-1)$$

$$C \text{ vs } A : c_2 = \bar{x}_1 - \bar{x}_3 = -1.18 ; (c_1=1 ; c_2=0 ; c_3=-1)$$

Test de Scheffe B vs C

$$C_{\text{critique}} = \sqrt{(k-1) F_{\alpha} S^2 r \sum_{i=1}^3 c_i^2 / n_i} = 0.64$$

$|C_1| < 0.64$  alors le contraste  $C_1$  n'est pas significativement différent de 0, on conclut que les moyennes  $\bar{x}_2$  et  $\bar{x}_3$  ne sont pas différentes.

$$C \text{ vs } A : C_{\text{critique}} = \sqrt{(k-1) F_{\alpha} S^2 r \sum_{i=1}^3 c_i^2 / n_i} = 0.83$$

$|C_2| > 0.83$  alors le contraste  $C_2$  est significativement différent de 0, on conclut que les moyennes  $\bar{x}_1$  et  $\bar{x}_3$  sont différentes.

### **Conclusion :**

Les traitements B et C ont la même efficacité et le traitement A est plus efficace que les traitements B et C .