Comparaison de plusieurs moyennes

L'analyse de variance (ANOVA) permet de comparer **les moyennes de plusieurs échantillons**. Les conditions préalables sont :

- les échantillons soient indépendants

- les distributions des mesures étudiées (les échantillons) soient issues de distributions parentes normales: *normalité des données*.

- les échantillons sont extraits de distributions parente de même variance (les variances observées sont homogènes): *homogénéité des variances*.

Remarque; ces conditions sont identiques à celles du test de Student dans la comparaison de 2 moyennes.

Si l'une de ces hypothèses n'est pas remplie, l'utilisation de l'ANOVA risquerait d'aboutir à des conditions erronées. L'exemple suivant est repris d'un cours consultable sur internet (http://www.univ-tours.fr/ash/psycho)

Analyse de variance à 1 facteur:

Supposons que l'on étudie les notes obtenues à une épreuve de math par 4 écoles différentes.

Groupes	A	B	C	D	
Notes	x_{A1} $\dots$	x_{B1} $\dots$	x_{C1} $\dots$ x_i $\dots$	x_{D1} $\dots$	
Effectif	n_A	n_B	n_C	n_D	$N = n_A + n_B + n_C + n_D$
Total	T_A	T_B	T_C	T_D	$T_G = T_A + T_B + T_C + T_D$
Moyenne	$\bar{x}_A$	$\bar{x}_B$	$\bar{x}_C$	$\bar{x}_D$	$\bar{x}_G = T_G / N$

Les trois conditions sont supposées vérifiées :
- « Les écoles sont indépendantes »
- La variable (note en math) se distribue normalement dans les ensembles parents des 4 classes.
- Les ensembles parents des 4 classes ont les mêmes variances.

Les hypothèses statistiques:

(H0) : moyennes identiques $\mu_A = \mu_B = \mu_C = \mu_D = \mu_G$

(H_1) : Au moins l'une des moyennes est différentes de μ_G

La solution repose sur la décomposition de la variation de la variable en une variation 'intra' groupe et une variation dite 'inter' groupe (variation ou somme des carrés).

<u>La variation totale:</u>

$$SC_T = \sum_{i=1}^{N} (x_i - \overline{x_G})^2$$

<u>La variation entre les classes</u> (variation inter-groupe ou factorielle)

$$SC_F = n_A(\overline{x_A} - \overline{x_G})^2 + \cdots + n_D(\overline{x_D} - \overline{x_G})^2$$

<u>La variation à l'intérieur de chaque classe</u> (variation intra-groupe ou résiduelle)

$$SC_r = \sum_{i=1}^{n_A} (x_{Ai} - \overline{x_A})^2 + \cdots + \sum_{i=1}^{n_D} (x_{Di} - \overline{x_D})^2$$

On en déduit les variances (ou carrés moyens):

$$CM_T = \frac{SC_T}{N-1}$$

$$CM_F = \frac{SC_F}{k-1}$$

$$CM_r = \frac{SC_r}{N-k}$$

Remarques:

$$SM_T = SM_F + SM_r$$

$$ddl_T = ddl_F + ddl_r$$

La statistique de décision:

Le rapport entre la variance inter-groupe et la variance intra-groupe suit une loi de Fisher Snédécor avec les degrés de liberté (ddl Inter, ddl Intra).

$$F = \frac{\text{variance inter}}{\text{variance intra}} = \frac{CM_F}{CM_r}$$

A partir des résultats que l'on observe sur les échantillons, on calcule une valeur de F et on compare cette valeur à une valeur critique (choisie avec un certain seuil, généralement 5% ou 1%).

- Si le f que l'on calcule est inférieur au f critique alors on est dans la zone de non rejet de H_0, le test est non significatif.

- Par contre si le f calculé est supérieur au f critique alors on est dans la zone de rejet de H_0. On accepte donc H_1 et on en conclut qu'une moyenne diffère des autres avec le risque d'erreur α. Il y a un effet du facteur étudié.

Source de variation	Somme des carrés des écarts	Nombre de ddl	Carrés moyens (variances)	F
Entre les groupes (inter)	$SC_F = \sum_{j=1}^{k}\left(\frac{T_j^2}{n_j}\right) - \frac{T_G^2}{N}$	k-1	$CM_F = \frac{SC_F}{k-1}$	$\frac{CM_F}{CM_r}$
A l'intérieur des groupes (intra)	$SC_r = \sum_{i=1}^{N} x_i^2 - \sum_{j=1}^{k}\left(\frac{T_j^2}{n_j}\right)$	N-k	$CM_r = \frac{SC_r}{N-k}$	
Total	$SC_T = \sum_{i=1}^{N} x_i^2 - \frac{T_G^2}{N}$	N-1		

Exemple : On reprend l'exemple précédent avec les notes suivantes

Groupes	A	B	C	D
Notes	6 3 7 5 4	8 8 5 6 7 6 2	7 4 8 6 5 9	4 3 6 3

Supposons que nous ayons vérifié les hypothèses de normalité et d'homogénéité des variances pour les 4 groupes.
Tester l'effet du facteur « écoles » sur les notes obtenues en math.

1/ La méthode utilisée sera bien entendu une méthode d'analyse de variance. Posez les hypothèses nulle et alternative à tester.

2/ Faites ensuite tous les calculs nécessaires :

Source de variation	Somme des carrés des écarts	Nombre de ddl	Carrés moyens (variances)	F
Entre les groupes (inter)				
A l'intérieur des groupes (intra)				
Total				

F de Snédécor avec 3 et 18 ddl :

p(F < 3,16) = 0,95 et p(F < 5,09) = 0,99

3/ Conclusion