Exemple 2 sur l'analyse de la variance (ANOVA)

L'analyse de la variance Dans cet exemple, on prend une série d'éléments représentée selon le tableau suivant :

$S1$	$S2$	$S3$
3	5	5
2	3	6
1	4	7

Pour faire l'ANOVA, il faut calculer la somme des carrés totale notée **SCT** . Pour cela il faut calculer la moyenne générale des trois séries données qui est $\bar{\bar{Y}}$.
$\bar{\bar{Y}} = \dfrac{3+2+1+5+3+4+5+6+7}{9} = 4.$
On peut aussi calculer $\bar{\bar{Y}}$ à partir des moyennes de chaque série comme suit :
Soient $\bar{Y_1}$, $\bar{Y_2}$ et $\bar{Y_3}$ respectivement les moyennes des séries $S1$, $S2$ et $S3$ tel que :
$\bar{Y_1} = \dfrac{3+2+1}{3} = 2.$
$\bar{Y_2} = \dfrac{5+3+4}{3} = 4.$
$\bar{Y_3} = \dfrac{5+6+7}{3} = 6.$

Donc :

$$\bar{\bar{Y}} = \frac{\bar{Y_1}+\bar{Y_2}+\bar{Y_3}}{3} = \frac{2+4+6}{3} = 4.$$

SCT $= (3-4)^2+(2-4)^2+(1-4)^2+(5-4)^2+(3-4)^2+(4-4)^2+(5-4)^2+(6-4)^2+(7-4)^2 = 30.$

Si on reprend le tableau vu précédemment, on remarque qu'on a 3 groupes de données on peut dire qu'on a m groupes de données et chaque groupe de données contient 3 élément, donc on peut dire que chaque groupe ou classe contient n données. En tout on m $\times$ n données. Si on connait la moyenne générale des données qui est $\bar{\bar{Y}}$ alors les m $\times$ n données sur le tableau ne sont pas toutes utiles ou indépendantes. C'est à dire si on connait les m $\times$ n-1 données, on peut utiliser la moyenne pour trouver la donnée manquante. On dit qu'on a m $\times$ n-1 degrés de liberté notée **DDL**. Dans notre cas le**DDL** $=$ 3$\times$3-1 $=$ 8.
Pour calculer la variance des données représentées dans le tableau précédent on utilise la formule suivante :

La variance $= \dfrac{SCT}{DDL} = \dfrac{30}{8}.$
La question qu'on peut poser maintenant est ce que cette variance est due à la variation au sein du groupe (à l'intérieur du groupe) ou entre les groupes ?
Pour cela on va décomposer la somme des carrés totale **SCT** en somme des carrés qui

provienent des variations à l'intérieur des classes et en somme des carrés qui proviennent des variations entre les classes.
On commence par calculer la somme des carrés intra-classes c'est à dire on calcule les écarts par rapport à la moyenne dans chaque classe cette somme est notée par SC_{intra}.

$SC_{intra} = (3-2)^2+(2-2)^2+(1-2)^2+(5-4)^2+(3-4)^2+(4-4)^2+(5-6)^2+(6-6)^2+(7-6)^2 = 2+2+2 = 6$.
$\Rightarrow SC_{intra} = 6$.
Maintenant si on prend par exemple la première classe S1, on remarque qu'on a on réalité deux données indépendantes car si on connait la moyenne de la classe qui est $\bar{Y_1}$ et deux autres éléments alors on pourra calculer l'autre élément. On général, si on a une classe qui contient n données et si on connait la moyenne, alors il y a n-1 données indépendantes. Cette valeur est notée par DDL $= m \times (n-1)$ sachant que m représente le nombre de classes dans lesquelles on a n-1 données indépendantes.
Conclusion : le degré de liberté à l'intérieur de la $SC_{intra} = m \times (n-1) = 6$.

maintenant, on va calculer la part de la variation totale qui est due à la variation entre les classes. Cette somme est la somme des carrés inter-classes notée par : SC_{inter}.

$SC_{inter} = (2-4)^2+(2-4)^2+(2-4)^2+(4-4)^2+(4-4)^2+(4-4)^2+(6-4)^2+(6-4)^2+(6-4)^2 = 12+0+12 = 24$, on remarque que la première classe contribue de 12 par rapport à la variation de la SC_{inter}qui est 24, la même chose pour la dernière classe.
On peut aussi se poser la même question c'est à dire : L'intervention des variables indépendante elle est de combien si on calcule la SC_{inter} (dans ce cas on cherche le degré de liberté) ? Dans ce cas, on se pose la question : si on connait $\bar{\bar{Y}}$ j'ai besoin de combien de moyenne de classe indépendantes ? La réponse est m-1 c'est à dire si je connais deux moyennes et la moyenne générale, alors je peux calculer la moyenne restante. Donc le DDl pour la SC_{inter} = m-1 = 2.

Remarque : SCT $= SC_{intra} + SC_{inter}$, la même chose avec les degrés de liberté c'est à dire $DDl_{SCT} = DDL_{SC_{intra}} + DDL_{SC_{inter}}$.

Interprétation des statistiques obtenus Si on prend l'exemple précédent, on peut considérer que cet exemple représente des groupes de patients. Chaque groupe on lui a donné un aliment différent. Dans ce cas chaque colonne représente le résultat obtenu sur chaque patient en lui donnant l'aliment selon son groupe.
La question qui se pose est : quel est l'impact des aliments ?
Si on revient sur le tableau vu précédemment, on remarque directement que $\bar{Y_3}$ est plus grand que les autres moyennes donc l'aliment donné au groupe 3 a vraiment un impact.
On peut aussi se poser une autre question : Est ce que cette différence des moyenne est due au hasard ou il y a vraiment un impact des aliments ? Autrement dit : Est ce que les moyennes obtenues sont identiques ou différentes des moyennes réelles ?
Si on note respectivement par μ_1, μ_2 et μ_3 les moyennes réelles des classes. Est ce que $\mu_1 = \mu_2 = \mu_3$. Alors, s'il y a une différence entre les moyennes, on peut dire que le type d'aliment qu'on donne a un impact. Pour cela on fait un test d'hypothèse.

On commence par la définition des hypothèses.
L'hypothèse nulle H_0 : L'aliment n'a pas d'impact $\Rightarrow \mu_1 = \mu_2 = \mu_3$.
L'hypothèse alternative H_1 : L'aliment a un impact $\Rightarrow \mu_1 \neq \mu_2 \neq \mu_3$.
newline On suppose que H_0 est vraie, puis on va calculer une statistique F qui va suivre une loi de Fisher, cette statistique est une caution de deux variables qui suivent une loi de Khi2.

F $= \dfrac{\dfrac{SC_{inter}}{m-1}}{\dfrac{SC_{intra}}{m \times (n-1)}}$, alors si $\dfrac{SC_{inter}}{m-1} >>> \dfrac{SC_{intra}}{m \times (n-1)}$, on peut dire que la variation due à des variations entre les classes est beaucoup plus importante que la variation due à des variations l'intérieur des classes. Si le F est grand, alors on a une très faible probabilité que l'hypothèse nulle H_0 soit vraie. Maintenant si F est petit c'est à dire $\dfrac{SC_{inter}}{m-1} <<< \dfrac{SC_{intra}}{m \times (n-1)}$ et dans ce cas on a une très grande probabilité que l'hypothèse nulle H_0 soit vraie.

Si on reprend l'exemple précédent F $= \dfrac{\frac{24}{2}}{\frac{6}{6}} = 12$.

Si on fixe le seuil de signification $\alpha = 0,1$ qui veut dire qu'on calcule la statistique F et si on a une probabilité inférieur à $0,1$ d'avoir obtenu une telle valeur de F alors on peut rejeter l'hypothèse nulle et si la probabilité d'une telle valeur de la statistique F est supérieur à $0,1$ alors on ne peut pas rejeter l'hypothèse nulle.
On prend la table de la loi de Fisher qui correspond à $\alpha = 0,1$ puis on prend l'intersection entre la colonne 2 qui correspond au $DDl_{SC_{inter}} = 2$ et la ligne 6 qui correspond au $DDl_{SC_{intra}} = 6$, cette intersection donne un F = 3,46 on peut aussi noter $F_\alpha = 3,46$ (appelée aussi F critique). Donc il y a une probabilité inférieur à (10 %) d'avoir une valeur supérieur à 3,46. Notre valeur de F appelé aussi $F_{2,6}$ selon les DDL est nettement supérieur à 3,46, donc on rejette H_0.

Conclusion : L'alimentation a vraiment un impact mais avec un risque $\alpha = 10$ % .