Université Larbi Ben M'Hidi –Oum El Bouaghi-

Faculté des sciences exactes et sciences de la nature et de la vie

Département des sciences exactes et sciences de la nature et de la vie

Module : Biostatistiques

Chapitre III : Analyse de la variance à un facteur (ANOVA à 1 facteur)

1. Définition :

L'analyse de la variance (**ANOVA)** est une généralisation du test de comparaison de deux moyennes en ce sens que le but est de comparer plusieurs groupes quand à leur moyennes.
C'est une technique qui trouve des utilisations dans plusieurs domaines comme la recherche biomédicale pour évaluer les effets d'un médicament en fonction de l'âge, l'agriculture pour évaluer l'importance du facteur génétique dans le rendement, etc.
Dans un modèle de l'analyse de la variance, on suppose une variable quantitative Y qu'on désigne sous le nom de variable dépendante ou variable expliquée et les variables qui définissent les groupes (variables qualitatives) sont les variables indépendante ou explicative (facteur). Cette terminologie provient du fait que le but est d'expliquer la variable Y en fonction d'un certain nombre de variables de groupement.

2. Types d'ANOVA :

On distingue trois types d'ANOVA : les types I, II et III. Le type I est qualifiée de modèle à effets fixes, les niveaux de chacun des facteurs étant déterminés délibérément par l'expérimentateur : c'est le cas dans la plupart des protocoles expérimentaux.
Le type II est appelée modèle à effets aléatoires, et dans ce type de modèle les niveaux du facteur d'étude sont déterminées de manière aléatoire. On s'intéresse alors principalement à la variabilité entre les échantillons par rapport à la variabilité à l'intérieur d'un échantillon. Enfin, le type III est un modèle à effets fixes et aléatoires, que l'on rencontre uniquement dans les ANOVA à plusieurs critères de classification ou facteurs

3. L'analyse de variance à un facteur :

Le choix du plan expérimental revêt toute son importance puisque ce sont le statut des variables et les relations qu'elles entretiennent entre elles qui guident la démarche d'analyse.

Pour les analyses de variance à un seul facteur, le facteur d'étude est le plus souvent déterminé par l'expérimentateur, et on se trouve dans une situation de type I. C'est le cas lorsque l'expérimentateur choisit délibérément les niveaux du facteur d'intérêt

4. Tableau de l'analyse de variance à un facteur :

	Critère de classification			
Données	Niveaux (groupes) 1	2	j	k
1	x_{11}	x_{12}	x_{1j}	x_{1k}
2	x_{21}	x_{22}	x_{2j}	x_{2k}
3	x_{31}	…		
…				
i	x_{i1}	…	x_{ij}	x_{ik}
…				
n_j	$x_{n_1 1}$	…	$x_{n_j j}$	$x_{n_k k}$
Totaux $\sum_{i=1}^{n_j} x_{ij}$	T_1	T_2	T_j	T_k
Effectifs n_j	n_1	n_2	n_j	n_k
Moyennes $\bar{x}_j$	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_j$	$\bar{x}_k$

Tableau : Distribution des observations dans un protocole à un seul facteur.

5. Principe de l'ANOVA :

Au cours de l'analyse de variance, on cherchera principalement à déterminer si les moyennes diffèrent entre elles dans leur globalité et lorsque c'est le cas, quelles sont les paires de moyennes qui sont significativement différentes.

6. Le facteur :

Le facteur est un caractère naturel (sexe, race, température, humidité,…etc.) ou artificiel construit, il permet de partager une population en catégorie socioprofessionnel, individus à traiter par un médicament et individus non traiter par ce médicament.

7. Décomposition de la somme des carrés :

7.1. Mesure de la variation totale (SCT) :

SCT = somme des carrés des écarts à la moyenne générale $\bar{Y}$, sans tenir compte du groupe (j = 1 … k) de provenance des données.

$$SCE_{totale} = \sum_{i=1}^{p} \sum_{j=1}^{n_j} (Y_{ij} - \bar{Y})^2$$

7.2. Mesure de la variation intragroupe (SCE) :

La variation [somme des (écarts par rapport à la moyenne)] à l'intérieur des groupes ne nous intéresse pas explicitement dans cette analyse. On considère qu'il s'agit de variation expérimentale.

Faisant la somme de ces termes pour tous les groupes j, on obtient

$$SCE_{intra} = \sum_{i=1}^{p} \sum_{j=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2$$

7.3. Mesure de la variation intergroupe (SCE) :

Pour chaque groupe *j*, il s'agit de calculer le carré de l'écart entre la moyenne de ce groupe et la moyenne générale, puis de sommer ces valeurs pour tous les groupes.

$$SCE_{inter} = \sum_{i=1}^{p} n_j (\bar{Y}_j - \bar{Y})^2$$

8. Hypothèses :

On teste l'hypothèse d'une absence de différence entre les k moyennes au niveau de la population parente : les k échantillons proviennent de la même population ou de populations ayant des caractéristiques comparables.

L'hypothèse nulle H_0 à tester est ainsi : H_0 : $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$,

H_1 l'hypothèse alternative étant que les échantillons sont issus de populations différentes : au moins deux des moyennes parentes diffèrent entre elles. H_1 : $\mu_1 \neq \mu_2$ ou $\mu_1 \neq \mu_3$ ou $\mu_2 \neq \mu_3$.

9. Test d'hypothèse :

Pour k échantillons d'effectifs n, on utilisera le test F de Fisher-Snedecor, qui est simplement le rapport entre le carré moyen des groupes (CMF) et le carré moyen de l'erreur (CMR) :

$F_{obs} = \frac{CMF}{CMR}$ qui est à comparer aux valeurs critiques de la distribution du F de Fisher-Snedecor au seuil α avec les degrés de liberté du numérateur et du dénominateur respectivement:
$v_1 = k - 1$ et $v_2 = n - k$.

10. Tableau de décomposition de la variance :

On présente généralement cette décomposition de la variance en un tableau résumant, pour les différentes sources de variation (groupe et erreur), les sommes des carrés des écarts à la moyenne (SC) ainsi que les degrés de liberté associés à chaque somme des carrés. Le carré moyen (CM) associés aux groupes (ou facteur) et à l'erreur, se retrouvent en faisant le rapport des somme des carrés sur leurs degrés de liberté respectifs.

Source de variation	Valeur	ddl	Carré moyen	F
Facteur	SCF	k - 1	CMF	CMF/CMR
Résiduelle	SCR	n – k	CMR= S^2	/
Total	SCT	n - 1	/	/

- **Exemple d'application :**

Le tableau suivant présente des mesures de la hauteur (en mm) de la plante *Saede brassica*, réalisées dans plusieurs milieux différents. Un chercheur désire comparer ces données afin de connaître l'effet du milieu sur la taille de *Saede brassica*.

Milieu 1	**Milieu 2**	**Milieu 3**	**Milieu 4**	**Milieu 5**
12	141	56	87	241
15	146	76	105	264
12	135	43	79	225
18	147	78	123	257

24	154	45	114	248
32		69		258
31				236
15				

La hauteur moyenne varie-t-elle selon les groupes ?

a) Hypothèses

H0 : toutes les moyennes sont égales

H1 : au moins une des moyennes est différente

Tableau de décomposition de la variance :

Source de variation	SC	ddl	Carré moyen	F
Facteur	217757.585	4	54439,396	341.345
Résiduelle	4146.608	26	159,485	/
Total	221904,195	30	/	/

c) Distribution de la variable :

Sous H_0, Fcritique suit une distribution de F à k – 1 = 4 et n - k = 31 - 5 = 26 degrés de liberté

d) Règle de décision :

À un seuil de 5%, la valeur critique du Fcritique (4, 26) ddl est de 2,74

H0 est acceptée si Fcritique < 2,74

e) Décision statistique :

F > Fcritique : H0 est rejetée

f) Conclusion biologique :

Au moins une des moyennes est différente : il y a un effet du milieu sur la hauteur des plantes.