9. Comparaison et analyse de variance

On souhaite étudier une variable aléatoire X dans k populations $P_1, \ldots, P_k$, par le biais de mesures sur des échantillons $E_1, \ldots, E_k$. La loi de X est supposée **gaussienne** (c'est-à-dire normale) dans chacune des populations. Le *test de Bartlett* permet de comparer les **variances** de X dans les différentes populations, tandis que l'*analyse de la variance* (à un facteur) permet de comparer les **moyennes** de X. Ces deux méthodes sont en général utilisées pour déterminer l'influence des différentes modalités d'**un** facteur sur le caractère étudié (par exemple l'effet de plusieurs médicaments sur un symptôme).

Pour chaque $i \in \{1, 2, \ldots, k\}$, on note μ_i et σ_i^2 l'espérance et la variance théoriques de X dans P_i, $\overline{x}_i$ et s_i^2 leurs estimations ponctuelles obtenues dans E_i ; n_i désigne la taille de E_i. On pose

$$n = \sum_{i=1}^{k} n_i$$

et on suppose les E_i indépendants.

On commence par la **comparaison des variances**. L'hypothèse nulle est :

$$(H_0) \ : \quad \sigma_1^2 = \cdots = \sigma_k^2 \ .$$

La *variance résiduelle*, ou *variance intragroupe* des échantillons est la moyenne des estimations des variances affectées des cœfficients $n_i - 1$:

$$s_R^2 = \frac{1}{n-k} \sum_{i=1}^{k} (n_i - 1) s_i^2 \ .$$

Cette variance caractérise la dispersion des valeurs **à l'intérieur** des échantillons. On note S_R^2 la variable aléatoire correspondante et $\lambda = 1 + \dfrac{1}{3(k-1)} \left(\displaystyle\sum_{i=1}^{k} \frac{1}{n_i - 1} - \frac{1}{n-k} \right)$, alors, sous l'hypothèse (H_0), la variable aléatoire

$$B = \frac{1}{\lambda} \Big((n-k) \ln(S_R^2) - \sum_{i=1}^{k} (n_i - 1) \ln(S_i^2) \Big)$$

suit sensiblement la loi du χ^2 à $k-1$ degrés de liberté. On calcule la valeur b prise par B sur les échantillons E_i ; on lit dans la table du χ^2 à $k-1$ degrés de liberté la valeur χ_α^2 telle que $P(B > \chi_\alpha^2) = \alpha$, puis :

- si $b \geq \chi_\alpha^2$, on rejette (H_0), avec risque d'erreur α ;
- si $b < \chi_\alpha^2$, on ne peut pas rejeter (H_0).

On continue avec la **comparaison des moyennes**, que l'on va traiter par la méthode d'*analyse de la variance*. Cette méthode suppose que X ait une loi gaussienne **de même variance** σ^2 dans toutes les populations. Le respect de ces conditions est d'autant plus important pour la validité du résultat que les effectifs des échantillons sont faibles et inégaux. L'hypothèse nulle est :

$$(H_0) \ : \quad \mu_1 = \cdots = \mu_k \ .$$

On note $\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i \overline{x}_i$ et s^2 les estimations ponctuelles de la moyenne et de la variance de X sur la réunion des échantillons. On définit la *variance factorielle*, ou *variance intergroupe* des échantillons par

$$s_F^2 = \frac{1}{k-1}\sum_{i=1}^{k} n_i(\overline{x}_i - \overline{x})^2 \ .$$

Cette variance caractérise la dispersion des valeurs **d'un échantillon à l'autre**, c'est-à-dire la variation **due à l'influence du facteur étudié**.

Le *théorème d'analyse de la variance* stipule que :

$$(n-1)s^2 = (n-k)s_R^2 + (k-1)s_F^2 \ ,$$

ce qui permet éventuellement d'obtenir la valeur de s_F^2 à partir de celles de s_R^2 et de s^2.

On note S_F^2 la variable aléatoire associée à la variance factorielle. Sous (H_0), la variable aléatoire $F = \frac{S_F^2}{S_R^2}$ suit la loi de Snédécor à $(k-1, n-k)$ degrés de liberté. On lit dans la table de Snédécor le nombre f_α tel que $P(F \geq f_\alpha) = \alpha$, puis :

- si $\frac{s_F^2}{s_R^2} \geq f_\alpha$, on rejette (H_0), avec risque d'erreur α ;
- si $\frac{s_F^2}{s_R^2} < f_\alpha$, on ne peut pas rejeter (H_0).

Exercice 1

On veut savoir si l'addition de substances adjuvantes à un vaccin modifie la production d'anticorps. Pour cela, on mesure les quantités d'anticorps produites par des sujets après administration de quantités égales du vaccin, additionné ou non d'une substance adjuvante. On obtient les taux :

— sans substance adjuvante : 1 ; 3 ; 3 ; 0 ; 1 ,

— avec de l'alumine : 2 ; 4 ; 5 ; 4 ; 3 ; 6 ,

— avec des sels de calcium : 3 ; 3 ; 4 ; 5 ,

— avec des phosphates : 1 ; 4 ; 2 ; 3 ; 3 .

(a) Quelles hypothèses faut-il faire pour pouvoir appliquer la technique d'analyse de la variance à la résolution du problème posé ? La validité de ces hypothèses est-elle importante dans le cas présent ?

(b) Sous les hypothèses adéquates, tester l'hypothèse selon laquelle les populations dont sont extraites les 4 échantillons ont la même variance.

(c) En précisant toujours les hypothèses adéquates, l'efficacité du vaccin dépend-elle :

(i) de la présence de substances adjuvantes ?

(ii) de leur nature ?

Exercice 2

On a étudié le développement d'un parasite à l'intérieur d'un organisme hôte, en fonction de la température d'élevage. La moyenne et l'écart-type estimés du nombre de jours de développement du parasite sont donnés dans le tableau suivant.

température (°C)	nombre d'animaux	moyenne et écart-type estimés
16	32	81 et 6,8
20	33	52 et 5,2
23	31	46 et 6,7

La température a-t-elle une influence sur la durée de développement du parasite ?

Exercice 3

On étudie l'activité d'un enzyme sérique, noté PDE, en fonction de différents facteurs dans l'espèce humaine. Les résultats sont exprimés en unités internationales par litres de sérum. On admettra que les populations considérées sont gaussiennes.

(a) Chez deux groupes de femmes, enceintes ou non, on obtient les résultats suivants :

enceintes : 4,2 ; 5,5 ; 4,6 ; 5,4 ; 3,9 ; 5,4 ; 2,7 ; 3,9 ; 4,1 ; 4,1 ; 4,6 ; 3,9 ; 3,5 ;
non enceintes : 1,5 ; 1,6 ; 1,4 ; 2,9 ; 2,2 ; 1,8 ; 2,7 ; 1,9 ; 2,2 ; 2,8 ; 2,1 ; 1,8 ; 3,7 ; 1,8 ; 3,1 .

La grossesse a-t-elle une influence significative sur l'activité de la PDE ?

(b) Afin d'évaluer la précocité de l'augmentation d'activité enzymatique lors de la grossesse, on pratique des dosages chez des femmes enceintes à différentes semaines d'aménorrhée. On obtient les résultats suivants (sur des échantillons indépendants) :

4 sem.	5 sem.	6 sem.	7 sem.	8 sem.
7,2	4,9	10,4	4,6	6,1
4,3	4,8	4,6	5,6	11,4
5,5	4,7	8,4	8,3	8,2
4,6	5,4	6,1	6,9	5,7
4,7	4,7	8,1	4,5	6,6
5,5	4,7	5,4	4,7	6,6
6,6	6,2	6,7	6,7	6,3
5,3	5,6	7,5	4,8	5,9
5,4	3,2	6,4	5,0	5,8
3,9	6,1	5,6	5,0	4,8
5,5	6,7	6,3	5,3	9,1
2,7	5,5	7,7	7,8	13,2

L'âge de la grossesse a-t-il une influence sur l'activité de l'enzyme ?