Ecole Supérieur de Technologie de Fkih Ben Salah

Biostatistique

III. Analyse de Variance

Génie Biologique & Industrie Agroalimentaire

2019-2020

M. MERZOUKI
m.merzouki@usms.ma

COMPARAISON DE PLUSIEURS MOYENNES

INTRODUCTION A L'ANALYSE DE VARIANCE ANOVA

L’hypothèse statistique

L'analyse de variance est une méthode statistique qui permet la comparaison des moyennes de plusieurs échantillons.

On veut comparer les valeurs moyennes de glycémie à jeun de quatre groupes de patients diabétiques qui reçoivent un traitement hypoglycémiant différent. Il s'agit de 4 traitements bien définis. Le facteur étudié est le traitement.

Soient : X la variable dont on veut comparer les moyennes (exemple :la glycémie).

• μ_1, μ_2, …., μ_k les moyennes vraies de X dans les k groupes que l'on veut comparer.

Les hypothèses testées s'écrivent :

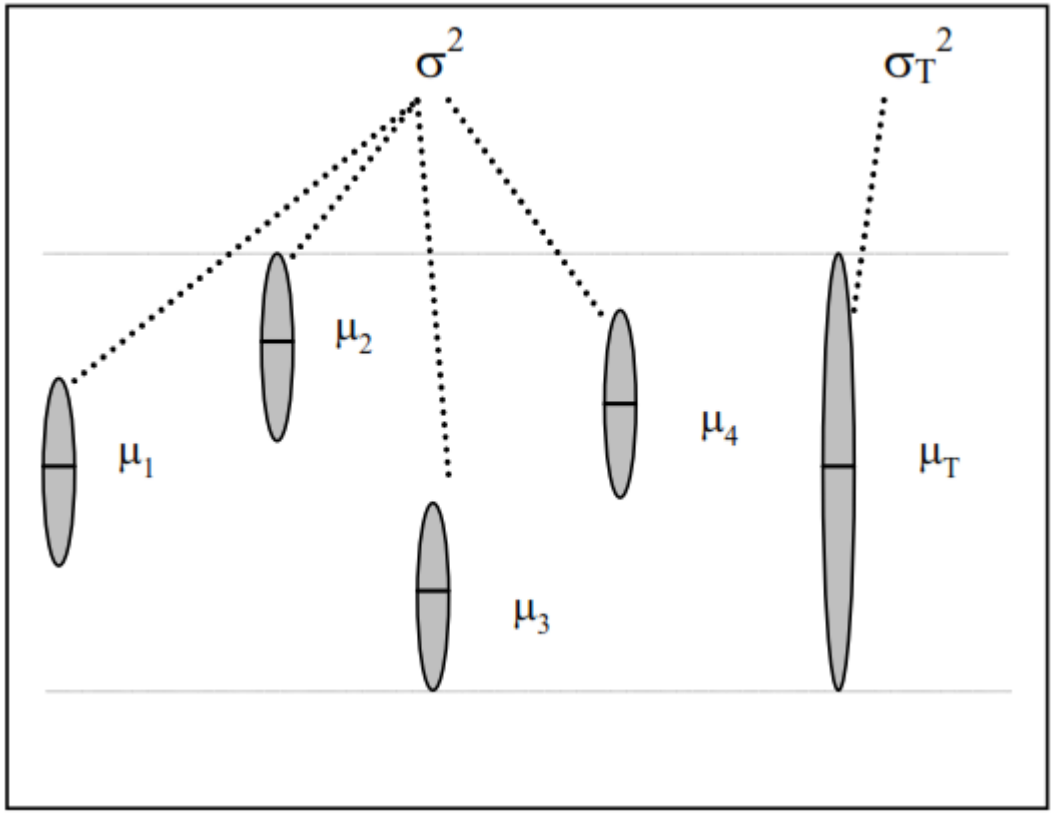
• H_0 : $\mu_1 = \mu_2 = \ldots = \mu$ ou les k groupes proviennent d'une population où X a une moyenne μ.

H_1: il y a au moins une moyenne différente entre les k moyennes, les k groupes ne proviennent pas de cette population.

Considérons le cas où il y a 4 populations différentes (k = 4) mais dans lesquelles la variance de X est la même : $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma^2$ (hypothèse nécessaire à la validité du test).

Si H_1 est vraie : Les moyennes μ_1, μ_2, μ_3 et μ_4 ne sont pas égales.

Si l'on regroupe les 4 populations, la moyenne générale est $\mu_T = (\mu_1 + \mu_{2+} \mu_{3+} \mu_4)/4$ et la variance totale est σ_T^2.

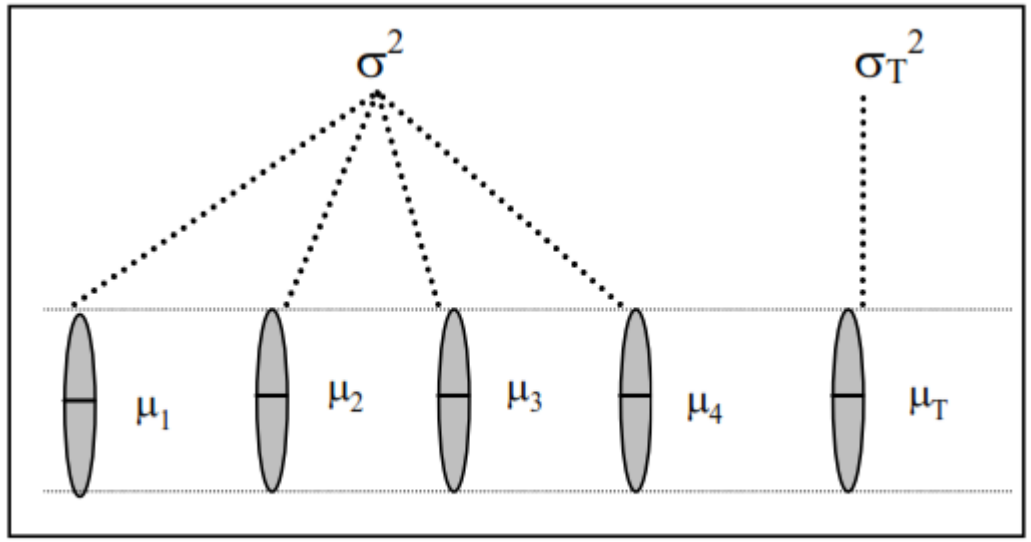


est σ_T^2 est d'autant plus grande que les moyennes μ_1, μ_2, μ_3 et μ_4 sont dispersées, c'est-à-dire que les différences entre ces 4 moyennes sont plus grandes.

les moyennes μ_1, μ_2, μ_3 et μ_4 sont différentes. La variance totale σ_T^2 est plus grande que la variance σ^2 de chacune des populations.

A l'inverse si H_0 est vraie :

Les moyennes μ_1, μ_2, μ_3 et μ_4 sont égales à μ, et σ_T^2 est égale à σ^2 qui est la variance au sein de chacune des populations



Les moyennes μ_1, μ_2, μ_3 et μ_4 sont égales. La variance totale σ_T^2 est égale à la variance σ^2 de chacune des populations.

On comprend donc qu'on puisse comparer les moyennes de X dans les différentes populations en comparant la variance σ^2 de X « à l'intérieur » de chacune des populations à la variance σ_T^2 de X obtenue en regroupant les populations.

L'ampleur de la dispersion totale (ou variabilité de x) σ_T^2 , dépend d'une part de l'ampleur de la dispersion au sein de chacune des k populations (variabilité intra population) mesurée par σ_T^2 et d'autre part de la dispersion entre ces populations (variabilité inter population).

L'analyse de variance consiste à comparer la variabilité intra population (due à des fluctuations d'échantillonnage) et la variabilité inter population (due à l'effet éventuel du traitement dans notre exemple

Si toutes les moyennes sont identiques la variance totale tend vers σ^2

Si elles ne sont pas identiques la variance totale tend vers $\sigma_T^2 > \sigma^2$

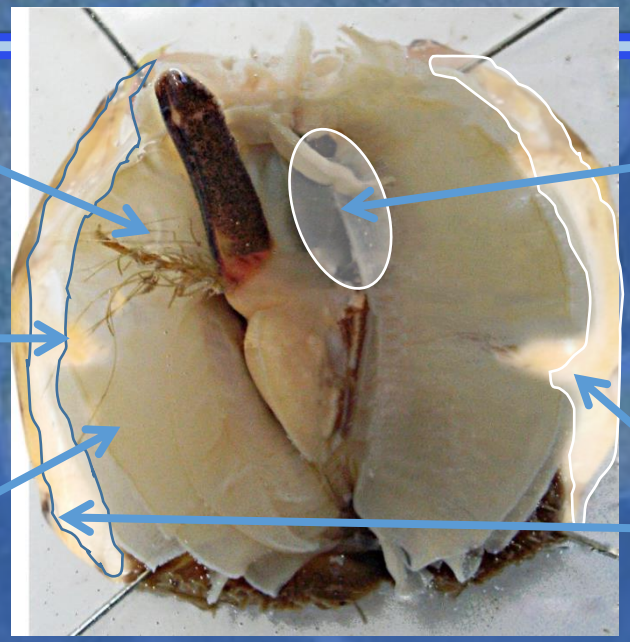
Les conditions d'application de l'analyse de variance imposent que la variable X ait une distribution normale et de même variance dans chacune des k populations.

Dans le cadre d'un travail de recherche en écotoxicologie, un dosage de Zn a été réalisé dans glandes digestives de moules collectées dans 3 zones différents d'une région côtière de l'Atlantique Marocain. Dans chaque zone, dix moules ont été prélevées au hasard, sur chacune desquelles un dosage de la teneur en zinc dans les G.D a été réalisé. On voudrait savoir si la teneur en zinc dans les G.D des moules varie de façon significative entre les trois zones. Les résultats sont donnés sous forme brute dans la table 1, en moyennes et écarts type estimés dans la table 2 et en diagrammes la figure 1.

✓ **Choix de l'espèce**

- **Reconnue mondialement comme étant une espèce bioindicatrice du milieu marin.**
- **Utilisée dans plusieurs programmes mondiaux de bio-surveillance marine.**
- **Modèle biologique étudié par notre équipe dans des travaux antérieurs et en cours.**

✓ **Dissection des animaux**




Zone 1
Zone 2
Zone 3

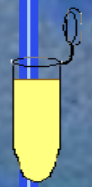
Bioaccumulation des métaux traces chez la moule *Mytilus galloprovincialis*

- ✓ **Dosage des métaux trace (Cd, Cu et Zn)**

manteau\ glande digestive

digestion acide à froid (HNO3 65%)

digestion acide à chaud (HNO3 65%)



l'ajout d' HNO3 a été fait dans un rapport de 1 ml d' HNO3 pour 0,5 g de poids

Spectromètre d'Absorption Atomique (SAA)

Moyenne ± ET (p.f)

Spectrophotomètre

Les éléments métalliques ont été dosés selon la méthode décrite par Amiard et al (1987)

Table 1: Teneurs en zinc dans les GD de moules collectées dans 3 zones

Z1	Z2	Z3
8,51	8,6	3,6
8,75	8,47	3,37
8,42	7,86	4,26
8,95	8,82	3,02
8,64	9,8	4,2
9,18	8,89	3,89
9,17	8,04	3,54
8,9	7,35	3,85
8,51	7,23	3,23
8,21	7,84	3,82

Table 2: Moyennes et écarts types des teneurs en zinc de moules collectées dans 3 zones

	Z1	Z2	Z3
μ_i	8,72	8,29	3,68
σ_i	0,32	0,78	0,40

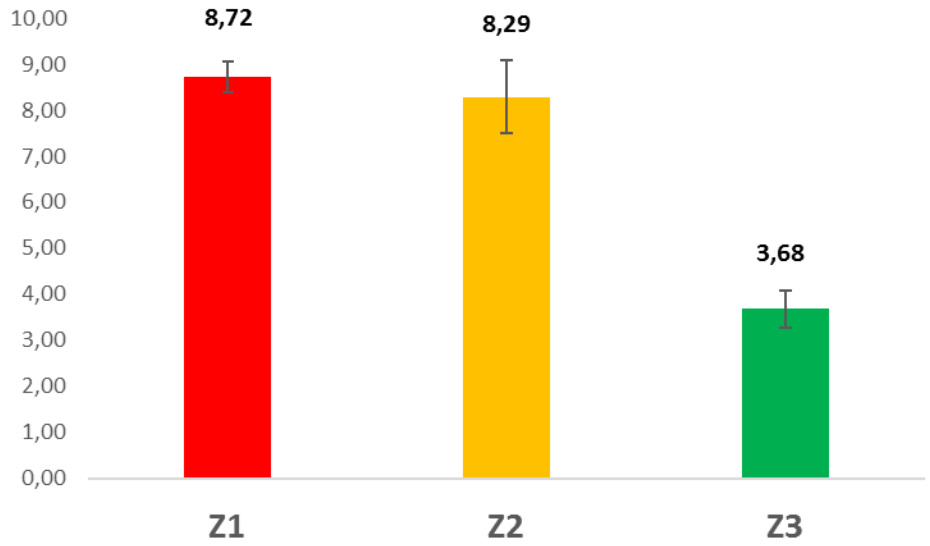


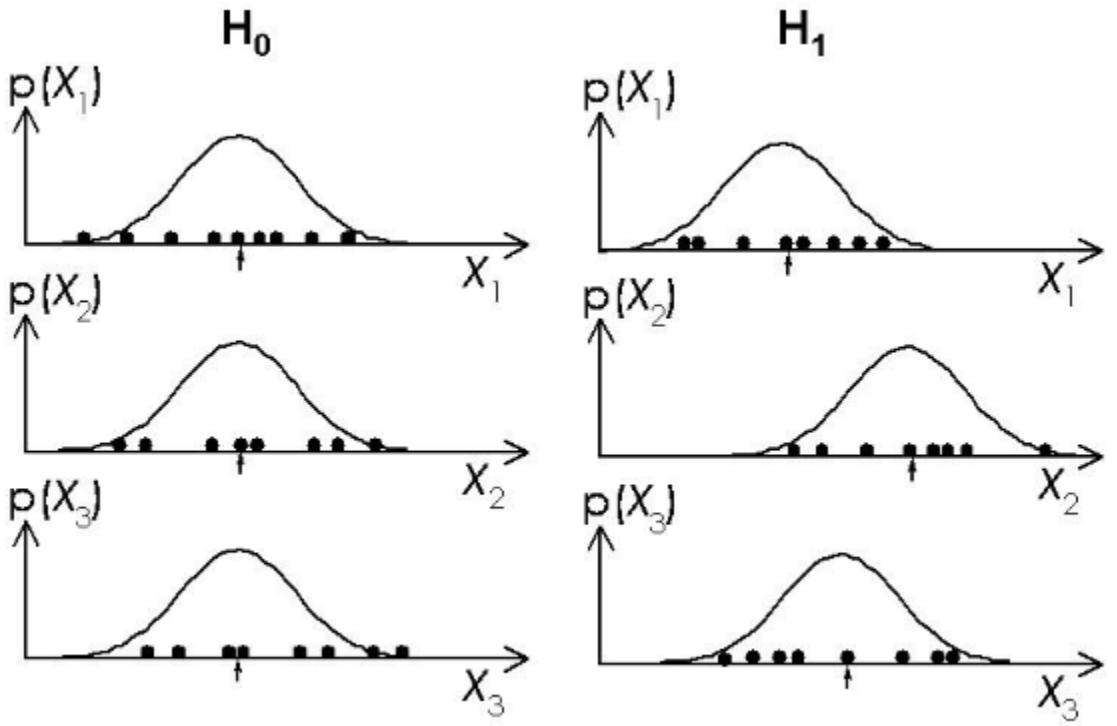
Figure 1: Teneurs en zinc de moules collectées dans 3 zones

Le facteur qualitatif étudié que l'on notera A est la zone étudiée et il peut prendre 3 modalités.

La variable quantitative continue observée que l'on notera X est la teneur en Zinc.

La variable X est observée sur 3 échantillons indépendants correspondant aux 3 modalités du facteur A

.



Le principe de l'analyse de variance est la décomposition de la variance ou de la variation de X d'où le nom de cette méthode. La variation ou Somme des Carrés des Ecarts totale (SCE_T) est décomposée en la somme de la variation ou Somme des Carrés des Ecarts factorielle ou inter-groupe (SCE_A) et de la variation ou Somme des Carrés des Ecarts résiduelle ou intra-groupe (SCE_R) selon les formules suivantes :

$$SCE_T = SCE_A + SCE_R$$

$$SCE_T = SCE_A + SCE_R$$

$$SCE_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij} - \overline{X}\right)^2 = N\,V(X)$$

$$F = \frac{CM_A}{CM_R}$$

$$SCE_A = \sum_{i=1}^{k} n_i\left(\overline{X}_i - \overline{X}\right)^2 = \left(\sum_{i=1}^{k} n_i \overline{X}_i^{\,2}\right) - N\overline{X}^2$$

$$SCE_R = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij} - \overline{X}_i\right)^2 = \sum_{i=1}^{k}(n_i - 1)\hat{\sigma}_i^2$$

k **représentant le nombre de modalités du facteur *A*, n_i représentant le nombre d'observations du groupe *n°i* et N =$\sum n_i$ le nombre total d'observations.**

A partir de ces variations on va estimer la variance ou carré moyen factoriel(le) ou inter-groupe CM_A et la variance ou carré moyen résiduel(le) ou intra-groupe CM_R par les formules suivantes :

$$CM_A = \frac{SCE_A}{k-1}$$

$$CM_R = \frac{SCE_R}{\sum_{i=1}^{k}(n_i - 1)} = \frac{SCE_R}{N-k}$$

Sous l'hypothèse H_0 la variance inter-groupe est en théorie égale à la variance intra-groupe puisqu'en fait tout se passe comme si les observations provenaient toutes du même groupe. Ainsi si la variance inter-groupe est nettement plus grande que la variance intra-groupe, cela veut dire que l'on n'est pas sous l'hypothèse H_0. En pratique la variable de décision utilisée est le rapport des deux variances $F=\frac{CM_A}{CM_R}$ qui suit la loi de Fisher et Snédécor de degrés de liberté $v_1 = k-1$ et $v_2 = N-k$ sous l'hypothèse nulle. Lorsque ce rapport est suffisamment grand on rejette l'hypothèse H. En pratique on estimera l'ordre de grandeur du degré de signification k à partir de la table unilatérale de la loi de Fisher et Snédécor.

***On calcule d'abord la* Somme des Carrés des Ecarts factorielle (*SCE_A*)**

	Z1	Z2	Z3
μ_i	8,72	8,29	3,68
σ_i	0,32	0,78	0,40

$$SCE_A = \sum_{i=1}^{k} n_i \left(\overline{X}_i - \overline{X}\right)^2 = \left(\sum_{i=1}^{k} n_i \overline{X}_i^{\,2}\right) - N\overline{X}^2$$

SCE_A = ((10×(8,72)²) + 10×(8,29)²) + 10×(3,68)²) – 30 × (6,89)²

SCE_A = 156,403

***Puis on calcule la* Somme des Carrés des Ecarts résiduelle (*SCE_R*)**

$$SCE_R = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_i\right)^2 = \sum_{i=1}^{k} (n_i - 1)\hat{\sigma}_i^2$$

SCE_R = ((9×(0,32)²) + 9×(0,78)²) + 9×(0,40)²)

SCE_R = 7,91

carré moyen factoriel(le) ou inter-groupe CM_A

$$CM_A = \frac{SCE_A}{k-1}$$

$$\mathbf{CM_A = \frac{156{,}403}{2} = 78{,}20}$$

	Z1	Z2	Z3
μ_i	8,72	8,29	3,68
σ_i	0,32	0,78	0,40

carré moyen résiduel(le) ou intra-groupe CM_R

$$CM_R = \frac{SCE_R}{\sum_{i=1}^{k}(n_i - 1)} = \frac{SCE_R}{N-k}$$

$$\mathbf{CM_R = \frac{7{,}91}{27} = 0{,}29}$$

266,900

Finalement on calcule la statistique $F = \frac{CM_A}{CM_R}$

$$F = \frac{78{,}20}{0{,}29} = 266{,}900$$

$$F \geq F_{(2,\ 27,\ 0,05)}$$

$$266{,}900 \leq 3{,}35$$

ce qui signifie que H0 doit être accépté. Il y a donc une déférence significative entre les zone. Par conséquent la quantité de Zn observée dépend de la zone