Analyse de la variance à un facteur ANOVA (1)

ANOVA(1) c'est quand?

- On utilise cette modélisation lorsque l'on souhaite expliquer une variable quantitative à l'aide d'une variable qualitative ayant k modalités

ANOVA(1) c’est quoi?

- L'analyse de la variance est la comparaison des moyennes de **k** populations, à partir d'échantillons aléatoires et indépendants prélevés dans chacune d'elles.
- Ces populations sont en général des variantes (ou niveaux **k**) d'un ou plusieurs facteurs contrôlés de variation (facteurs **A**, **B**, ...).

Conditions d'applications de l'ANOVA

- les populations étudiées suivent une distribution normale
- les variances des populations sont toutes égales (**HOMOSCEDASTICITE**)
- les échantillons E_i de tailles n_i sont prélevés aléatoirement et indépendamment dans les populations.

Procédure de calcul d'une ANOVA

- Déterminer si les échantillons varient de la même manière.
- Si nous démontrons l'homogénéité des variances, alors nous pouvons comparer les moyennes de ces échantillons.

Problèmes liés à l'égalité des variances

Test de l'homogénéité des variances

- $(H_0)_\sigma$: les variances sont homogènes
- $(H_1)_\sigma$: Au moins une des variances est différente des autres

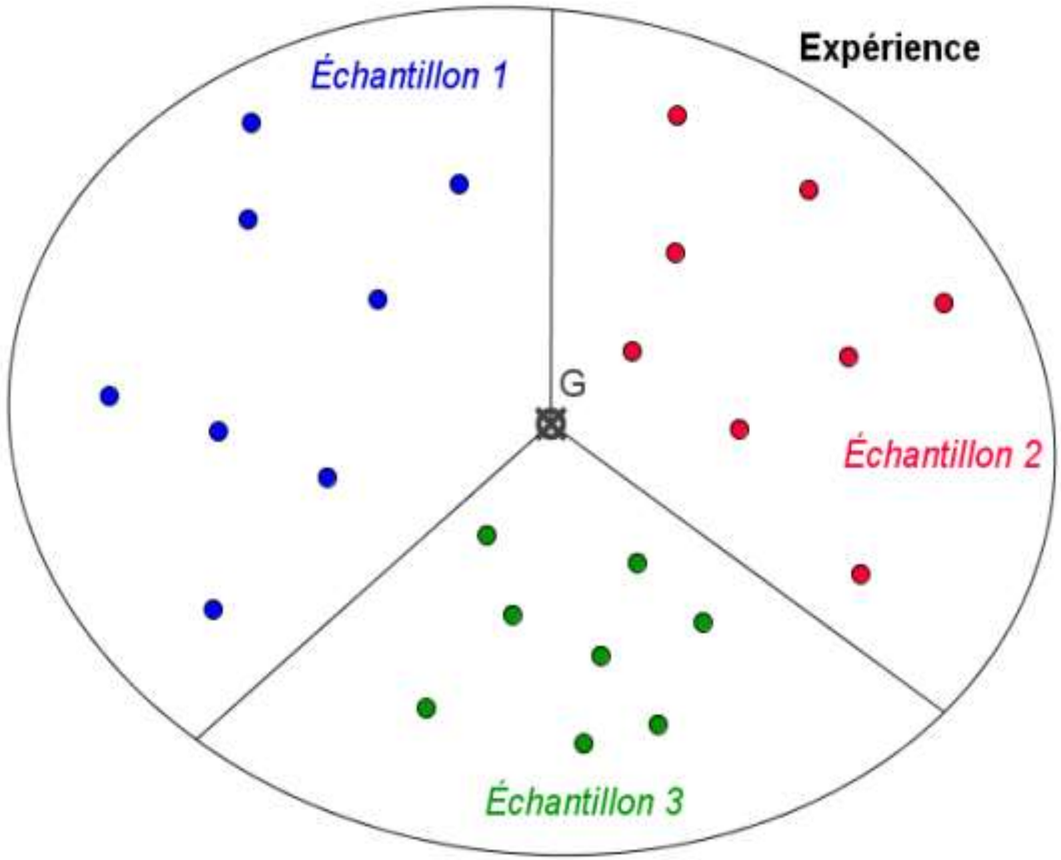
→ utilisation d'un test de comparaison de plusieurs variances

Conclusion

- Si $(H_0)_\sigma$ est rejetée : il est théoriquement impossible de comparer des échantillons qui ne varient pas de la même manière.
- Si $(H_0)_\sigma$ n’est pas rejetée : par conséquent, il est possible de comparer les moyennes de tels échantillons

Schématisation de l'analyse multiple de moyennes

- Expérience avec k échantillons - Données initiales



Comparaison des moyennes - Hypothèses

- H_0 : toutes les moyennes sont identiques)
- H_1 : au moins une des moyennes est différente des autres

Autrement dit

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \\ H_1 : \exists i, j \in \{1, \dots, k\} \quad \mu_i \neq \mu_j \end{cases}$$

Variances totale, factorielle, résiduelle

Pour chaque échantillon E_i de taille n_i , on calcule :

- Moyenne

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

- variance expérimentale

$$S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_i \right)^2$$

Pour l'ensemble de l'expérience :

- Taille totale

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

- Moyenne générale

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij} = \frac{1}{n}\sum_{i=1}^{k} n_i \bar{x}_i$$

- Variance totale

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{X})^2$$

- Variance factorielle (variance intergroupe) :

c'est la dispersion des valeurs d'un échantillon à l'autre (influence du facteur) :

$$S_F^2 = \frac{1}{k-1}\sum_{i=1}^{k}(\bar{x}_i - \bar{X})^2$$

- Variance résiduelle (variance intragroupe)

C'est la dispersion des valeurs à l'intérieur des échantillons (variabilité individuelle) :

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

- Variance factorielle et variance résiduelle : estimation de la variance de la population (σ^2)

Théorème d'analyse de la variance

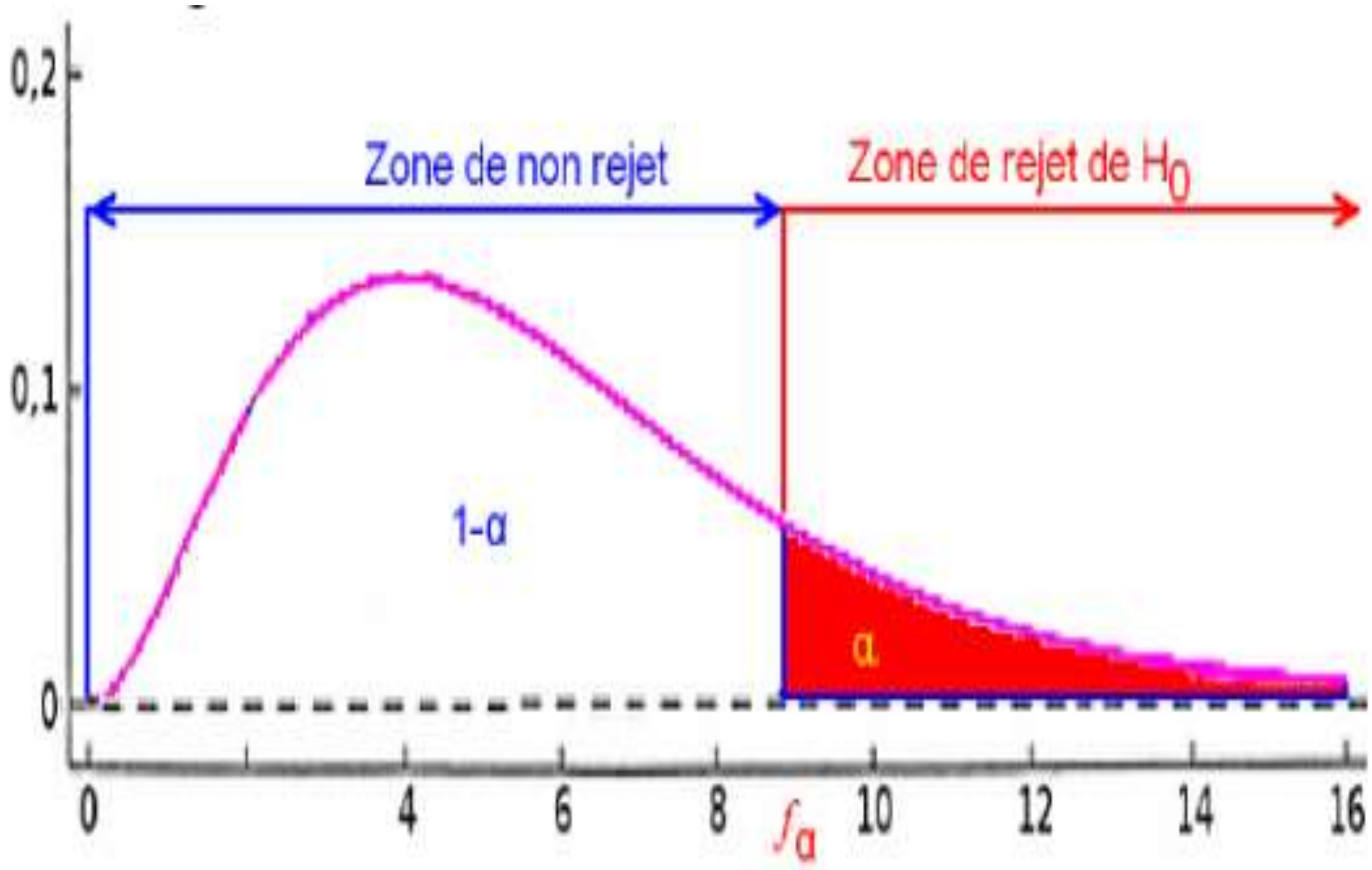
Variation	SCE	ddl	CM	F
Factorielle	$SCE_F = (k-1)\cdot s_F^2$	k-1	$CM_F = s_F^2$	$F = \frac{s_F^2}{s_R^2}$
Résiduelle	$SCE_R = (n-k)\cdot s_R^2$	n-k	$CM_R = s_R^2$	
Totale	$SCE_T = (n-1)\cdot s^2$	n-1	$CM_T = s^2$	

Avec $SCE_F + SCE_R = SCE\ _T$

Sous l'hypothèse H_0 :

- F suit une loi de Snédécor à $v_1 = k - 1$ et $v_2 = n - k$ ddl
- (test unilatéral : le rapport n'est pas obligatoirement supérieur à 1)
- Choix du risque: risque de première espèce α (erreur commise lorsqu'on rejette à tort).

Décision



Zone de non rejet
Zone de rejet de H_0
0,2
0,1
0
1-α
α
0
4
6
8
f_α
10
12
14
16

Décision

- Si $F > f_{\propto}$ => **rejet de** H_0 **au risque** $\propto$:
  - La variance factorielle est significativement supérieure à la variance résiduelle : les moyennes diffèrent significativement entre-elles. Donc on attribue une influence significative au facteur étudié.
  - Recherche du degré de signification p (recherche du risque $\propto$ le plus petit possible pour conclure au rejet de H_0)
- Sinon rien ne permet de dire que les moyennes des populations ne sont pas égales => H_0 **n'est pas rejetée**.