# Laïus 5 : BIO 2042
# Test d'homoscédasticité et plans d'expérience

Guillaume Blanchet - Département de sciences biologiques, Université de Montréal – Janvier 2007

**Objectifs :**

A) Apprendre à tester l'homogénéité de variance
1) Test de Bartlett
2) Test de Levene

B) Apprendre à construire un plan d'expérience.

## A) Test d'homoscédasticité

Pour beaucoup de tests paramétriques (p.ex. ANOVA), l'homogénéité des variances (homoscédasticité) est une condition nécessaire. Plusieurs méthodes (plus de 80 !) existent pour tester l'homogénéité des variances dans plusieurs groupes qui n'ont pas nécessairement le même nombre d'objets.

Après une revue de la littérature, des simulations et encore quelques tests très préliminaires, nous avons remarqué que certains tests étaient plus appropriés selon la situation d'utilisation :

- Si l'échantillon est **petit** ($n \leq 50$) le **test de Bartlett** semble plus puissant
- Si l'échantillon est **grand** ($n > 50$) le **test de Levene** semble plus robuste

# 1) Test d'homoscédasticité de Bartlett

**Objectif :**

Tester l'équivariance entre plusieurs groupes (en particulier sur de petits jeux de données)

**Calculs :**

On utilise une statistique $B_c$

$$B_C = B / C$$

$$B = 2,3026\left[\left(\log_{10} s_{xp}^2\right)\underbrace{\sum_{i=1}^{k}\left(n_i - 1\right)}_{N-k} - \sum_{i=1}^{k}\left(n_i - 1\right)\log_{10} s_{x_i}^2\right]$$

$$C = 1 + \frac{1}{3\left(k-1\right)}\left[\sum_{i=1}^{k}\frac{1}{\left(n_i - 1\right)} - \frac{1}{\sum_{i=1}^{k}\left(n_i - 1\right)}\right]$$

$$s_{xp}^2 = \frac{\sum_{i=1}^{k}\left(n_i - 1\right)s_{x_i}^2}{\sum_{i=1}^{k}\left(n_i - 1\right)}$$

$s_{x_i}^2$ = Variance du groupe *i*

$k$ = Nombre de groupes

$n_i$ = Nombre d'éléments dans le groupe *i*

*N* = nombre total d'éléments

On compare la statistique $\boldsymbol{B_c}$ à une statistique $\chi^2$ à $\boldsymbol{k - 1}$ degrés de liberté.

## Exemple : Nombre de sapins poussant dans 3 forêts différentes.

| | Forêt 1 | Forêt 2 | Forêt 3 |
|---|---|---|---|
| | 45 | 78 | 354 |
| | 34 | 69 | 338 |
| | 35 | 86 | 351 |
| | 29 | 58 | 332 |
| | 42 | 57 | 341 |
| | 37 | 64 | 358 |
| | 44 | | 347 |
| | 28 | | |
| Variance ($S_x^2$) → | 42,214 | 131,867 | 86,476 |

**Test en 9 étapes**

**<u>Étape 1 :</u> Question biologique**

Les variances des trois groupes sont-elles égales ?

**<u>Étape 2 :</u> Déclaration des hypothèses**

$H_0$ : Les variances des trois groupes sont égales.

$$\sigma_1 = \sigma_2 = \sigma_3$$

$H_1$ : Au moins une des variances diffère des autres.

$$\sigma_i \neq \sigma_j \text{ pour au moins un } i \neq j$$

**<u>Étape 3 :</u> Choix du test**

Le test utilisé est un test de Bartlett où :

$$B_C = \frac{B}{C}$$

où $B = 2,3026\left[\left(\log s_{xp}^2\right)\sum_{i=1}^{k}\left(n_i - 1\right) - \sum_{i=1}^{k}\left(n_i - 1\right)\log s_{x_i}^2\right]$

et $C = 1 + \frac{1}{3\left(k-1\right)}\left[\sum_{i=1}^{k}\frac{1}{\left(n_i - 1\right)} - \frac{1}{\sum_{i=1}^{k}\left(n_i - 1\right)}\right]$

et $s_{xp}^2 = \frac{\sum_{i=1}^{k}\left(n_i - 1\right)s_{x_i}^2}{\sum_{i=1}^{k}\left(n_i - 1\right)}$

et $s_{x_i}^2$, variance du groupe *i*.

et *k* = Nombre de groupes

et $n_i$ = Nombre d'éléments dans le groupe *i*

**<u>Étape 4 :</u> Condition d'application**

- Normalité des distributions

**<u>Étape 5 :</u> Distribution de la variable auxiliaire**

Si $H_0$ est vraie, la variable $B_c$ suivra une distribution du $\chi^2$ à $k - 1 = 3 - 1 = 2$ degrés de liberté.

**<u>Étape 6 :</u> Règles de décision**

On rejette $H_0$ au seuil $\alpha = 0,05$ si $B_c > \chi^2_{(0,05;\,2)} = 5,99$. (Table V du Scherrer)

**<u>Étape 7 :</u> Calcul du test**

$$s_{xp}^2 = \frac{(8-1)\times 42,214 + (6-1)\times 131,867 + (7-1)\times 86,476}{(8-1)+(6-1)+(7-1)} = 81,8717$$

$$B = 2,3026\left[\left(\log(81,8717)\right)\times(7+5+6) - \left(7\times\log(42,21429) + 5\times\log(131,8667) + 6\times\log(86,47619)\right)\right] = 1,92529$$

$$C = 1 + \frac{1}{3\cdot(3-1)}\left(\left(\frac{1}{7}+\frac{1}{5}+\frac{1}{6}\right) - \frac{1}{7+5+6}\right) = 1,075661$$

$$B_C = \frac{1,92529}{1,075661} = 1,789866$$

**<u>Étape 8 :</u> Décision statistique**

Comme $B_c$ = 1,789866 < 5,99, on ne peut pas rejeter $H_0$ au seuil $\alpha$ = 0,05.

**<u>Étape 9 :</u> Interprétation biologique**

Les trois échantillons de sapin ont la même variance. (Il n'y a pas forcément d'interprétation biologique à ce test, qui est souvent réalisé surtout pour des raisons techniques).

## 2) Test d'homoscédasticité de Levene

**Objectif :**

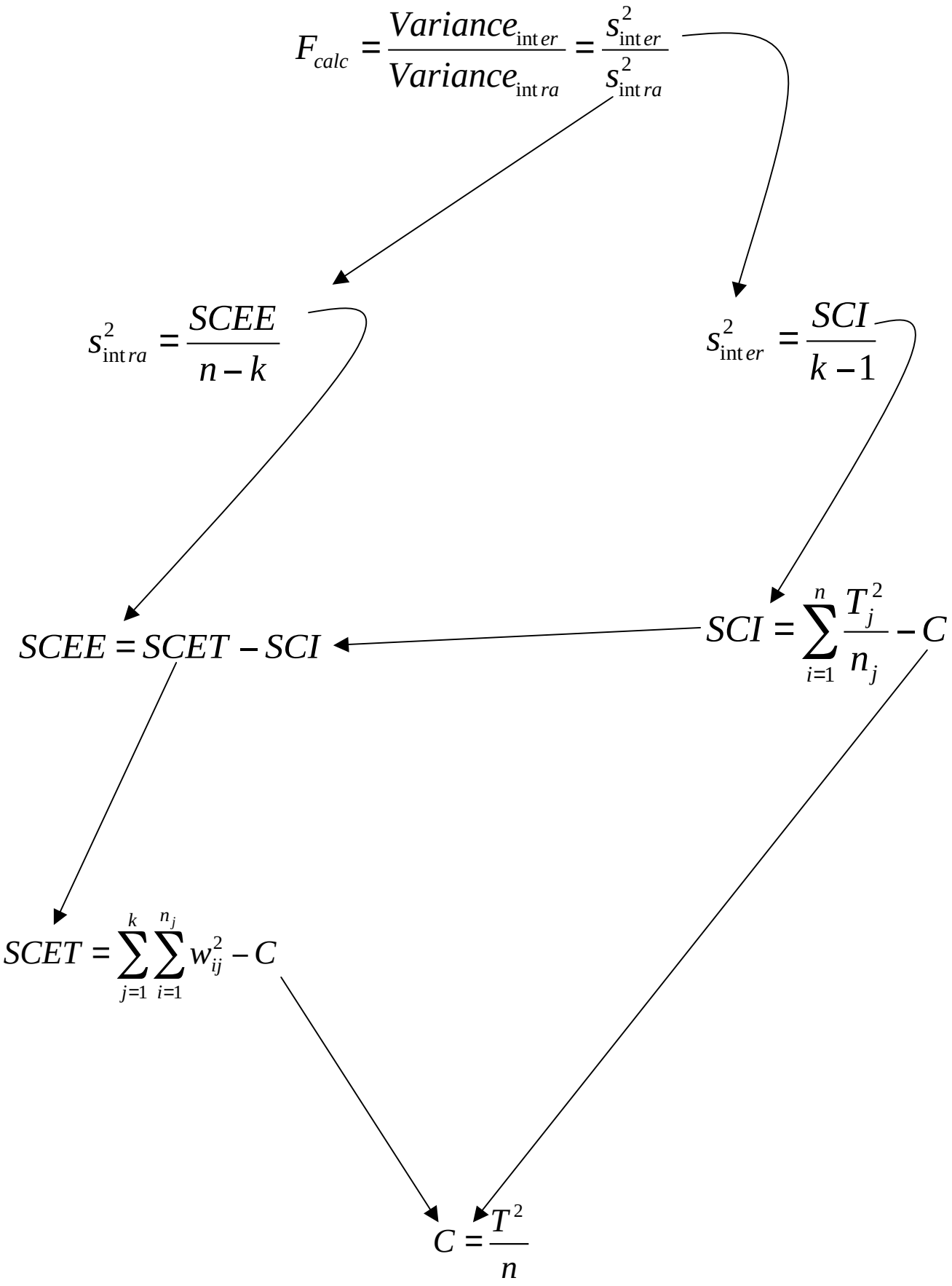
Tester l'équivariance entre plusieurs groupes (surtout pour les grands jeux de données)

**Calculs :**

| | **Groupe** | | |
|---|---|---|---|
| | 1 | … | *j* |
| Donnée 1 | $x_{11}$ | | $x_{1j}$ |
| … | | | |
| Donnée i | $x_{i1}$ | | $x_{ij}$ |

$$w_{ij} = \left| x_{ij} - Me_j \right|$$

| | **Groupe** | | | | |
|---|---|---|---|---|---|
| | 1 | … | *j* | … | *k* |
| Donnée 1 | $w_{11}$ | | $w_{1j}$ | | $w_{1k}$ |
| … | | | | | |
| Donnée i | $w_{i1}$ | | $w_{ij}$ | | $w_{ik}$ |
| Totaux | $\boldsymbol{T_1}$ | $\boldsymbol{T_{...}}$ | $\boldsymbol{Tj}$ | | $T = \sum T_j$ |
| Effectif | $\boldsymbol{n_1}$ | $\boldsymbol{n...}$ | $\boldsymbol{n_j}$ | | $n = \sum n_j$ |


$F_{calc} = \frac{Variance_{inter}}{Variance_{intra}} = \frac{s^2_{inter}}{s^2_{intra}}$
$s^2_{intra} = \frac{SCEE}{n-k}$
$s^2_{inter} = \frac{SCI}{k-1}$
$SCEE = SCET - SCI$
$SCI = \sum_{i=1}^{n} \frac{T_j^2}{n_j} - C$
$SCET = \sum_{j=1}^{k} \sum_{i=1}^{n_j} w_{ij}^2 - C$
$C = \frac{T^2}{n}$

## Exemple : Nombre de sapins poussant dans 3 forêts différentes.

| | Forêt 1 | Forêt 2 | Forêt 3 |
|---|---|---|---|
| | 45 | 78 | 354 |
| | 34 | 69 | 338 |
| | 35 | 86 | 351 |
| | 29 | 58 | 332 |
| | 42 | 57 | 341 |
| | 37 | 64 | 358 |
| | 44 | | 347 |
| | 28 | | |
| Mediane($Me_j$) ⟶ | 36 | 66,5 | 347 |

**Test en 9 étapes ...**

**<u>Étape 3 :</u> Choix du test**

Le test utilisé est un test d'homogénéité de variance de Levene :

$$F_{calc} = \frac{Variance_{inter}}{Variance_{intra}} = \frac{s^2_{inter}}{s^2_{intra}}$$

**...**

**<u>Étape 5 :</u> Distribution de la variable auxiliaire**

Si $H_0$ est vraie, la statistique $F_{calc}$ suivra une distribution du Fisher-Snedecor à $\nu_1 = k - 1 = 2$ degrés de liberté et $\nu_2 = n - k = 18$

**<u>Étape 6 :</u> Règles de décision**

On rejette $H_0$ au seuil $\alpha$ = 0,05 si $F_{calc} > F_{(0,05;2,18)}$ = 3,55.

**<u>Étape 7 :</u> Calcul du test**

$$C = \frac{T^2}{n} = \frac{21904}{21} = 1043,048$$

$$SCET = \sum_{j=1}^{k}\sum_{i=1}^{n_j} w_{ij}^2 - C = 1515,5 - 1043,048 = 472,4524$$

$$SCI = \sum_{i=1}^{n} \frac{T_j^2}{n_j} - C = 1092,786 - 1043,048 = 49,7381$$

$$SCEE = SCET - SCI = 472,4524 - 49,7381 = 422,7143$$

$$s^2_{intra} = \frac{SCEE}{n-k} = \frac{422,7143}{21-3} = 23,48413$$

$$s^2_{inter} = \frac{SCI}{k-1} = \frac{49,7381}{3-1} = 24,86905$$

$$F_{calc} = \frac{s^2_{inter}}{s^2_{intra}} = \frac{24,86905}{23,48413} = 1,058973$$

**<u>Étape 8 :</u> Décision statistique**

Comme $F_{calc}$ = 1,058973 < 3,55, on ne peut pas rejeter $H_0$ au seuil $\alpha$ = 0,05.

**<u>Étape 9 :</u> Interprétation biologique**

Les trois échantillons de sapin ont la même variance.

# B) Plan d'expérience

## Mise en situation 1

### Problématique générale

Une espèce nuisible de chenille vit dans les arbres feuillus des zones tempérées. Dans le but d'éliminer ces chenilles, on décide de tester l'effet d'un nouvel insecticide dans une petite forêt.

### Question

Le nombre de chenilles est-il influencé par la présence du nouvel insecticide ?

### Hypothèses

$H_0$ : la présence de l'insecticide n'influence pas le nombre de chenilles.
$H_1$ : la présence de l'insecticide influence le nombre de chenilles.

### Définition des composantes

**Facteur** (mécanisme agissant sur un système donné, et que l'on veut isoler et faire varier)

**Variable dépendante**

**Élément**

**Population statistique**

**Facteurs externes** (Facteur pouvant interférer avec le facteur étudié…Bref, source de confusion)

**Plan d'expérience**

# Mise en situation 2

## Problématique générale

Un nouveau cépage de la vigne a été créé pour survivre aux conditions climatiques du Québec, le seul problème est qu'il produit peu de raisin. On souhaite tester l'influence qu'auraient quatre fertilisants sur la production en raisin du nouveau cépage.

## Question

La production en raisin du nouveau cépage est-il influencé par la présence de fertilisants ?

## Hypothèses

$H_0$ : la présence de fertilisant n'influence pas la production en raisin du nouveau cépage.
$H_1$ : la présence de fertilisant influence la production en raisin du nouveau cépage.

## Définition des composantes

**Facteur** (mécanisme agissant sur un système donné, et que l'on veut isoler et faire varier)

**Variable dépendante**

**Élément**

**Population statistique**

**Facteurs externes**

**Plan d'expérience**