

Chapitre 9

Analyse de variance

Sommaire

1. Introduction.....	2
2. Conditions d'application.....	2
2.1. Structure des données.....	3
2.2. Conditions d'application.....	4
<i>2.2.1. Indépendance.....</i>	4
<i>2.2.2. Normalité.....</i>	4
<i>2.2.3. Homoscédasticité.....</i>	4
<i>2.2.4. Robustesse.....</i>	5
3. Modèle de l'analyse de variance	6
3.1. Modèle sous H_0 : homogénéité des données.....	6
3.2. Modèle sous H_1 : hétérogénéité des données.....	6
3.3. Equation fondamentale de l'analyse de variance.....	7
<i>3.3.1. Estimation des paramètres du modèle</i>	7
<i>3.3.2. Décomposition de la variation totale.....</i>	7
<i>3.3.3. Le rapport de corrélation.....</i>	8
4. Pratique de l'analyse de variance	9
4.1. Principe du test.....	9
4.2. Application et Tableau de variation.....	10

1 Introduction

Dans le cadre des tests d'hypothèses, nous avons émis des hypothèses concernant la moyenne d'une population (test de conformité) puis comparé les moyennes de deux populations (test d'homogénéité). Ce chapitre a trait à la comparaison des moyennes de plusieurs populations (> 2). **L'analyse de variance** peut être vue comme une **comparaison multiple de moyennes**. Dans tous les cas, la variable étudiée est un caractère quantitatif de type continu qui suit une loi normale.

Il existe différentes types d'analyse de variance qui se distinguent par le **nombre de facteurs** étudiés (un facteur, deux facteurs, deux facteurs avec répétitions, etc), la **nature du facteur** (caractère qualitatif ou quantitatif) et la **nature des modalités** associées au facteur (modèle fixe, modèle aléatoire, modèle mixte).

Nous ne traiterons que le cas de l'analyse de variance à un facteur contrôlé (modèle fixe) et ceci pour deux types de données. Dans ce cas **les conclusions ne s'appliquent qu'aux p modalités étudiées**.

◆ Les différentes modalités du facteur correspondent aux différentes modalités d'un caractère qualitatif. Ces **modalités sont déterminées** par l'expérimentateur.

Exemple : On étudie l'effet de différentes types d'alimentation (Paille, Foin, Herbe, Aliment ensilé) sur le rendement des vaches laitières.

◆ Les modalités du facteur correspondent à différentes valeurs prises par une **variable aléatoire dont les valeurs sont fixées par l'expérimentateur**. L'analyse de variance à un facteur peut également permettre d'étudier la relation, pas forcément linéaire, entre deux variables quantitatives. C'est le test de linéarité.

Exemple : On étudie l'effet de différentes doses d'engrais (20, 30, 40, 50) sur le rendement du blé. Pour chaque dose d'engrais expérimentée, le rendement est étudié sur un échantillon de parcelles donné.

2 Conditions d'application

L'analyse de variance à un facteur contrôlé ou ANOVA1 a pour objectif de **tester l'effet d'un facteur A** sur une **variable aléatoire continue**. Ceci revient à comparer les moyennes de plusieurs populations normales et de même variance à partir d'échantillons aléatoires et indépendants les uns des autres. Chaque échantillon est soumis ou correspond à une modalité du facteur A. Le terme ANOVA indique que la comparaison multiple de moyennes correspond en fait à la comparaison de deux variances.

2.1 Structure des données

Les données relatives à une analyse de variance à un facteur contrôlé sont structurées dans un tableau du type suivant :

<i>Facteur A</i>					
<i>modalité 1</i>			<i>modalité i</i>		<i>modalité p</i>
y_{11}			y_{i1}		y_{p1}
.			.		.
y_{1j}			y_{ij}		y_{pj}
.			.		.
y_{1n_i}			.		.
			y_{in_i}		y_{pnp}
\bar{y}_1			\bar{y}_i		\bar{y}_p

Notation :

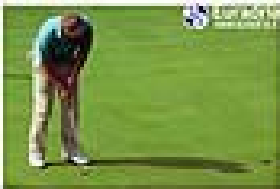
Le facteur contrôlé A présente p **modalités** ($1 \leq i \leq p$). On parle aussi de niveaux ou traitements.

Le nombre de **répétitions j** pour une **modalité i** est noté n_i . Le nombre de répétitions pour chaque modalité du facteur n'est pas forcément le même.

La valeur prise par la variable aléatoire Y pour la **modalité i** du facteur et **la répétition j** est notée y_{ij} et les valeurs moyennes pour chaque modalité notée \bar{y}_i .

Exemple :

International	National	Régional	"Récréational"
24,8	45,6	33,4	31,1
26,7	41,1	34,6	35,7
27,5	34,3	36,4	37,3
30,6	37,6	39,1	39,4
32,4	39,5	43,8	40,4
38,2		47,9	44,5
40,5		49,9	45,4
42,9		51,2	49,8
			50,1



Un test psychologique a été passé par 30 sportifs évoluant à des niveaux de compétition différents : international, national, régional et « récréational ». Une des mesures réalisées porte sur l'anxiété des sportifs au moment de la compétition. Celle-ci diffère-t-elle en fonction du niveau de compétition ? Réponse.

2.2 Conditions d'application de l'analyse

Les hypothèses relatives au modèle d'analyse de variances sont nombreuses. L'analyse des **résidus e_{ij}** ($y_{ij} - \bar{y}_i$) est particulièrement utile pour répondre aux hypothèses de normalité et d'homoscédasticité. Mais dans le cadre d'un modèle à effet fixe, il est équivalent de tester ces hypothèses sur la variable y_{ij} .

2.2.1 Indépendance

L'indépendance entre les différentes valeurs de la variable mesurée y_{ij} est **une condition essentielle** à la réalisation de l'analyse de variance.

Les p échantillons comparés sont **indépendants**.
L'ensemble des N individus est réparti au hasard (**randomisation**) entre les p modalités du facteur contrôlé A , n_i individus recevant le traitement i .

2.2.2 Normalité

La variable quantitative étudiée suit **une loi normale** dans les p populations comparées.
La variable aléatoire étudiée Y dont y_{ij} est une représentation, suit une loi normale $\mathcal{N}(\mu_i, \sigma)$ sous H_1 .

La normalité de la variable pourra être testée à l'aide d'un **test de Khi-deux d'ajustement** si les effectifs sont suffisamment importants. Sinon le test non paramétrique de **Lilliefors** permet de tester l'ajustement à loi normale lorsque les effectifs sont faibles.

On peut en fait se limiter à vérifier si la **distribution** des valeurs e_{ij} ou y_{ij} est **unimodale**.

Remarque : Si la normalité de la variable n'est pas vérifiée, soit on peut transformer cette dernière, soit avoir recours à l'équivalent non paramétrique de l'ANOVA, le **test de Kruskal-Wallis**.

2.2.3 Homoscédasticité

Les p populations comparées ont **même variance**.
Le facteur A agit seulement sur la moyenne de la variable Y et ne change pas sa variance.

Différents tests permettent de vérifier l'égalité des variances relatives aux p populations comparées

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_i^2 = \dots = \sigma_p^2 = \sigma^2 .$$

- Le **test de Lévène** est le test **le plus satisfaisant** pour effectuer la comparaison multiple de variances mais sa réalisation est assez longue car il correspond à une ANOVA1 sur les résidus e_{ij} .
- Le **test de Bartlett** est dédié à la comparaison multiple de variances avec **un nombre de répétitions n_i différent** selon les modalités i du facteur. Mais ce test est très sensible à l'hypothèse de normalité des p populations (peu robuste).
- Le **test de Hartley** est dédié à la comparaison multiple de variances avec un **nombre de répétitions n_i identiques** selon les modalités i du facteur. Mais ce test est très sensible à l'hypothèse de normalité des p populations (peu robuste).

Remarque : Si l'hétérogénéité entre variances est très importantes, on peut avoir recours aux statistiques non paramétriques, **test de Kruskal-Wallis**.



Exemple : Dans le cadre de l'exemple des sportifs, les conditions d'application de l'analyse de variance sont-elles vérifiées ? **Réponse.**

2.2.4 Robustesse

De nombreux travaux ont étudié la **robustesse** de l'ANOVA1 vis à vis des écarts aux hypothèses faites.

Hypothèses	Test	Robustesse
Normalité de Y	Test du χ^2 d'ajustement	Très robuste si indépendance et égalité des variances
Homoscédasticité des p distributions	Test de Levène ou de Bartlett	Très robuste à l'inégalité des variances
Indépendance des p distributions	Plan expérimental	Pas robuste

Remarque : L'analyse de variance à un facteur contrôlé est relativement peu sensible à l'inégalité des variances ainsi qu'à la non normalité lorsque les échantillons comparés sont de grandes tailles.

3 Modèle de l'analyse de variance

3.1 Modèle sous H_0 : homogénéité des données

L'analyse de variance à un facteur teste l'effet d'un facteur contrôlé A ayant p modalités sur les moyennes d'une variable quantitative Y .

L'hypothèse nulle testée est la suivante :
 il n'y a pas d'effet du facteur A et les p moyennes sont égales à une même moyenne μ .
 $H_0 : \mu_1 = \mu_2 = \dots = \mu_i = \dots = \mu_p = \mu$
 alors $y_{ij} = \mu + e_{ij}$ sous H_0
 avec e_{ij} variables aléatoires **indépendantes** suivant une **même loi normale $\mathcal{N}(0, \sigma)$** .

Les résidus e_{ij} correspondent aux fluctuations expérimentales pour chaque valeur de la variable y_{ij} mesurée.

Sous l'hypothèse $e_{ij} \rightarrow \mathcal{N}(0, \sigma)$, on montre que $y_{ij} \rightarrow \mathcal{N}(\mu, \sigma)$

$$\begin{aligned} E(y_{ij}) &= E(\mu + e_{ij}) = E(\mu) + E(e_{ij}) = \mu + E(e_{ij}) = \mu && \text{puisque } E(e_{ij}) = 0 \\ V(y_{ij}) &= V(\mu + e_{ij}) = V(\mu) + V(e_{ij}) = 0 + \sigma^2 = \sigma^2 && \text{puisque } V(e_{ij}) = \sigma^2 \end{aligned}$$

3.2 Modèle sous H_1 : hétérogénéité des données

L'hypothèse alternative est la suivante :
 il y a un effet du facteur A et il existe au moins deux moyennes significativement différentes.
 $H_1 : \exists \mu_i \neq \mu_j$
 alors $y_{ij} = \mu + a_i + e_{ij}$ sous H_1
 avec e_{ij} : variables aléatoires **indépendantes** suivant **une même loi normale $\mathcal{N}(0, \sigma)$** .
 a_i : **l'effet de la modalité i** du facteur A sur la variable Y

Sous l'hypothèse $e_{ij} \rightarrow \mathcal{N}(0, \sigma)$, on montre que $y_{ij} \rightarrow \mathcal{N}(\mu + a_i, \sigma)$

$$\begin{aligned} E(y_{ij}) &= E(\mu + a_i + e_{ij}) = E(\mu) + E(a_i) + E(e_{ij}) = \mu + a_i + E(e_{ij}) = \mu + a_i && \text{puisque } E(e_{ij}) = 0 \\ V(y_{ij}) &= V(\mu + a_i + e_{ij}) = V(\mu) + V(a_i) + V(e_{ij}) = 0 + 0 + \sigma^2 = \sigma^2 && \text{puisque } V(e_{ij}) = \sigma^2 \end{aligned}$$

Ainsi il existe une différence entre les moyennes de la variable selon les modalités du facteur contrôlé.

Remarque : Tester l'hypothèse nulle « absence d'effet sur facteur A » revient à tester H_0 : les a_i sont tous nuls.

3.3 Equation fondamentale de l'analyse de variance

3.3.1 Estimation des paramètres des modèles

Sous H_0 : $y_{ij} = \mu + e_{ij}$

$$\hat{\mu} = \bar{y} \quad \text{notée aussi } y_{..} \quad \bar{y} = \frac{1}{N} \sum_{i=1}^p \sum_{j=1}^{n_i} y_{ij} \quad \text{avec } N = \sum_{i=1}^p n_i$$

L'ensemble des données du tableau peuvent être estimées à partir de la moyenne totale des y_{ij} à laquelle s'ajoute la part d'aléatoire dans les mesures, e_{ij} .

Sous H_1 : $y_{ij} = \mu + a_i + e_{ij}$

$$\hat{\mu} + \hat{a}_i = \bar{y}_i \quad \text{notée aussi } y_{i.} \quad \bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \quad \text{moyenne des } y_{ij} \text{ pour la modalité } i$$

d'où $\hat{a}_i = \bar{y}_i - \bar{y}$

et $\hat{e}_{ij} = y_{ij} - \hat{\mu} - \hat{a}_i = y_{ij} - \bar{y} - \bar{y}_i + \bar{y} = y_{ij} - \bar{y}_i$

ainsi $\hat{e}_{ij} = y_{ij} - \bar{y}_i$

3.3.2 Décomposition de la variation totale

Soit le modèle sous H_1 : $y_{ij} = \mu + a_i + e_{ij}$

avec les estimateurs $y_{ij} = \bar{y} + (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i)$

$$(y_{ij} - \bar{y}) = (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i)$$

avec les écarts au carré $(y_{ij} - \bar{y})^2 = (\bar{y}_i - \bar{y})^2 + (y_{ij} - \bar{y}_i)^2 + 2(\bar{y}_i - \bar{y})(y_{ij} - \bar{y}_i)$

avec les sommes pour tous les individus j

$$\sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = n_i (\bar{y}_i - \bar{y})^2 + \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + 2(\bar{y}_i - \bar{y}) \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)$$

or $2(\bar{y}_i - \bar{y}) \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i) = 0$

car $\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i) = 0$ sachant que $E(e_{ij}) = 0$ si les hypothèses initiales sont vérifiées.

avec la somme pour les p modalités du facteur contrôlé

$$\sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^p n_i (\bar{y}_i - \bar{y})^2 + \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

L'équation fondamentale de l'analyse de variance

$$\sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^p n_i (\bar{y}_i - \bar{y})^2 + \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

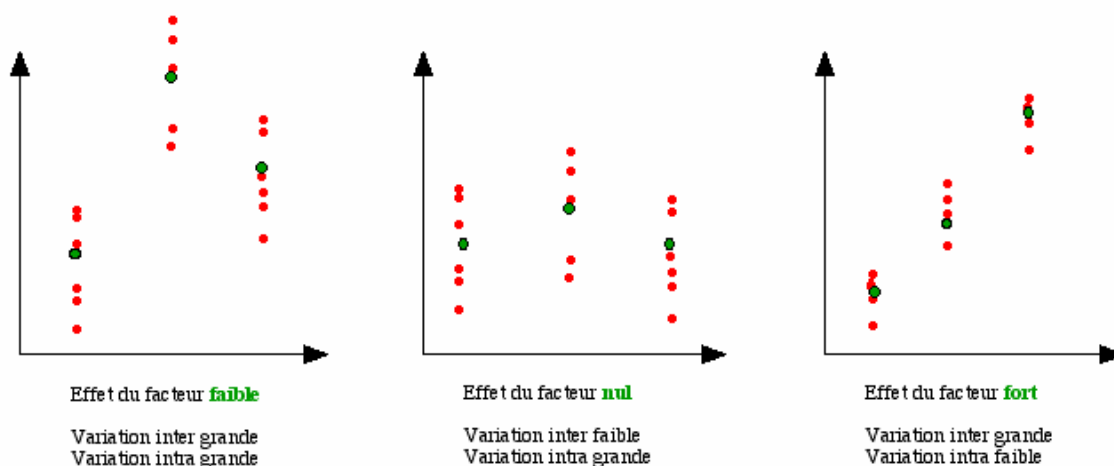
SCE_{totale}
SCE_{inter}
SCE_{intra}

Notation :

SCE_{totale} = somme des écarts totaux ou variation totale = $N s_y^2$

SCE_{inter} = somme des écarts liés aux effets du facteur A ou variation inter (entre modalités)

SCE_{intra} = somme des écarts résiduelles ou variation intra (interne à chaque modalité)

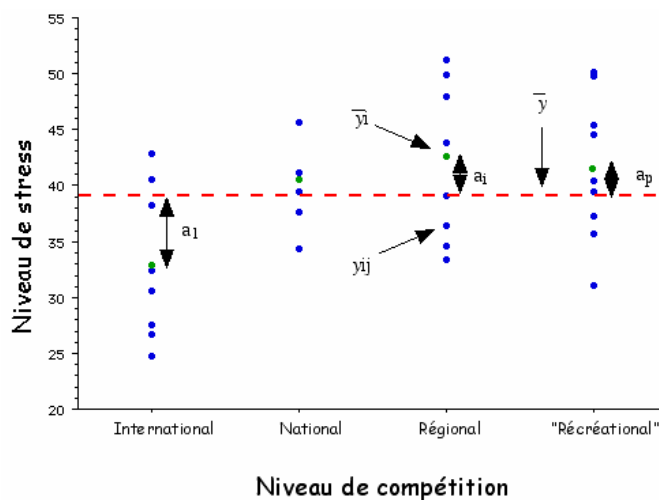


Exemple : Distribution des valeurs de la variable « anxiété des sportifs » y_{ij} en fonction des niveaux de compétition pour 30 sportifs.

\bar{y}_i : **moyenne** pour chaque modalité i du facteur **en vert**,

\bar{y} : **moyenne** de l'ensemble des données

a_i : effet de la modalité i sur la variable y_{ij}



3.3.3 Le rapport de corrélation

Le **rapport de corrélation** donne la part de la variabilité totale des données expliquée par l'effet du facteur A :

$$\eta^2 = \frac{SCE_{inter}}{SCE_{totale}}$$

C'est un **indice de liaison**, pas nécessairement linéaire entre les variables étudiées qui varie entre 0 et 1. Il mesure **la qualité de l'ajustement** des effets du facteur au travers des moyennes.

4 Pratique de l'analyse de variance

L'analyse de variance à un facteur teste l'effet d'un facteur contrôlé A ayant p modalités sur les moyennes d'une variable quantitative Y .

L'hypothèse nulle testée est la suivante

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_1 = \dots = \mu_p = \mu \quad \text{contre} \quad H_1 : \exists \mu_i \neq \mu_j$$

Le test de comparaison multiple de moyenne revient à faire un test de [comparaison de deux variance](#).

4.1 Principe du test

Soit l'équation de décomposition de la variation totale :

$$\sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^p n_i (\bar{y}_i - \bar{y})^2 + \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

SCE_{totale} SCE_{inter} SCE_{intra}

Les estimations des variances associées ou **carré moyen** sont :

Variance totale :	$\frac{SCE_{totale}}{N-1}$	avec $N = \sum_{i=1}^p n_i$
Variance due au facteur A (CM_{inter}) :	$\frac{SCE_{inter}}{p-1}$	estimateur de σ^2 si H_0 vraie
Variance résiduelle (CM_{intra}) :	$\frac{SCE_{intra}}{N-p}$	estimateur de σ^2 quelque soit le modèle

Remarque : L'équation fondamentale de l'analyse de variance ne s'applique pas aux variances : Variance totale \neq Variance inter + Variance intra

◆ **Sous H_0 :** $\mu_1 = \mu_2 = \dots = \mu_1 = \dots = \mu_p = \mu$ avec $a_i = 0 \quad \forall i$

$$\frac{SCE_{inter}}{p-1} \text{ et } \frac{SCE_{intra}}{N-p} \quad \text{estimateur du même paramètre } \sigma^2$$

d'où $\frac{SCE_{inter}}{p-1} \approx \frac{SCE_{intra}}{N-p}$

et donc le rapport $\frac{SCE_{inter}}{p-1} \times \frac{N-p}{SCE_{intra}}$ est **proche de 1**

◆ **Sous H_1 :** $\exists \mu_i \neq \mu_j$ avec au moins un $a_i \neq 0$

$$\frac{SCE_{intra}}{N-p} \quad \text{unique estimateur de } \sigma^2$$

d'où $\frac{SCE_{inter}}{p-1} \neq \frac{SCE_{intra}}{N-p}$ avec $\frac{SCE_{inter}}{p-1} \gg \frac{SCE_{intra}}{N-p}$

et donc le rapport $\frac{SCE_{inter}}{p-1} \times \frac{N-p}{SCE_{intra}}$ est **très supérieur à 1**

Sous H_0 : il n'y a pas d'effet du facteur A sur la variable Y
 $\mu_1 = \mu_2 = \dots = \mu_i = \dots = \mu_p = \mu$
 contre H_1 : le facteur A exerce un effet en moyenne sur la variable Y
 $\exists \mu_i \neq \mu_j$

$$F_{obs.} = \frac{CM_{inter}}{CM_{intra}} = \frac{\frac{SCE_{inter}}{p-1}}{\frac{SCE_{intra}}{N-p}}$$

suit une **loi de Fisher-Snedecor**

F_{obs} comparée à F_{seuil} lue dans **la table de la loi de Fisher-Snedecor** pour un **risque d'erreur α** fixé et $(p-1, N-p)$ degrés de liberté.

- si $F_{obs} > F_{seuil}$ l'hypothèse H_0 est rejetée au risque d'erreur α : le facteur contrôlé A a un effet significatif en moyenne sur les valeurs de la variable étudiée.
- si $F_{obs} \leq F_{seuil}$ l'hypothèse H_0 est acceptée: le facteur contrôlé A n'a pas d'effet significatif en moyenne sur les valeurs de la variable étudiée.

Remarque : Le rejet de l'égalité des moyennes ne permet pas de savoir quelles sont les moyennes significativement différentes. Pour cela, la méthode des contrastes ou **méthode de Scheffé** associée à l'analyse de variance permet de répondre à cette question.

4.2 Application et Tableau de variation

Le tableau de variation donne un résumé des calculs effectués pour l'analyse de variance.

Sources de variation	Degrés de liberté	Somme des Carrés des Ecarts	Carré Moyen	Test de Fisher-Snedecor
Totale	$N - 1$	SCE_{TOT}		$F_{obs.} = \frac{CM_{inter}}{CM_{intra}}$
Facteur	$p - 1$	SCE_{inter}	$CM_{inter} = SCE_{inter} / (p-1)$	
Résiduelle	$N - p$	SCE_{intra}	$CM_{intra} = SCE_{intra} / (N-p)$	

Pour effectuer les calculs, des **formules développées** peuvent être utilisées.

$$SCE_{TOT} = \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^p \sum_{j=1}^{n_i} y_{ij}^2 - \frac{T^2}{N} \quad \text{avec } T = \sum_{i=1}^p \sum_{j=1}^{n_i} y_{ij} \text{ et } N = \sum_{i=1}^p n_i$$

$$SCE_{inter} = \sum_{i=1}^p n_i (\bar{y}_i - \bar{y})^2 = \sum_{i=1}^p \frac{T_i^2}{n_i} - \frac{T^2}{N} \quad \text{avec } T_i = \sum_{j=1}^{n_i} y_{ij}$$

$$SCE_{\text{intra}} = \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = \sum_{i=1}^p \sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^p \frac{T_i^2}{n_i} \quad \text{ou} \quad SCE_{\text{intra}} = SCE_{\text{TOT}} - SCE_{\text{inter}}$$

**Exemple :**

L'anxiété des sportifs au moment de la compétition diffère-t-elle en fonction du niveau de compétition ? ([Tableau de variation](#))