Cours 2 : Les tests paramétriques

1. Généralités sur les tests

Un test d'hypothèse consiste à définir une règle de décision concernent la validité d'une hypothèse portant sur un modèle statistique au vu du résultat de l'cervation.

On définit en premier lieu une hypothèse H_0, dite hypothèse nulle, et son alternative H_1, dite hypothèse alternative.

.L'hypothèse nulle H_0 joue un rôle particulier, elle prétendra que les différences cervés entre valeurs calculées et valeurs théoriques sont dues au hasard. Si on doit rejeter l'hypothèse nulle H_0, on dira que les écarts cervés sont significatifs et on choisira l'hypothèse alternative H_1.

Un test statistique est alors une procédure qui permet, sur la base de l'cervation, d'accepter ou de rejeter l'hypothèse nulle H_0 qui est la seule hypothèse testée et celle qui permet les calculs conduisant à la conclusion.

Si les hypothèses testées portent sur une fonction du paramètre du modèle statistique à valeurs dans $\mathbb{R}^k$, on parle de "test paramétrique", dans les autres cas on parle de "test non paramétrique".

Si on suppose que la connaissance du paramètre d'un modèle statistique permet de savoir si H_0 est vraie ou fausse, on peut alors toujours se ramener au cas où les hypothèses à tester sont deux parties disjointes de l'espace des paramètres. Dans ce cas, on parlera d'"hypothèse simple" si l'hypothèse est réduite à un point de l'espace des paramètres, sinon on parlera d'"hypothèse composites".

Définitions 1

1. La statistique du test *: c'est une fonction qui résume l'information sur l'échantillon (ou valeur) qu'on veut tester. On la choisit de façon à pouvoir*

calculer sa loi sous H_0.

2. La région critique *: c'est la région de rejet de l'hypothèse nulle H_0 : on rejette H_0 si la valeur cervée de la statistique calculée à partir des données, appartient à la région de rejet.*

3. Test bilatéral *: c'est lorsque la région critique est partagée en deux parties, et dans ce cas le test prend la forme suivante :* $\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$

La région critique est alors de la forme : $]-\infty, a]U[b, +\infty[$.

4. Test unilatéral *: c'est lorsque la région critique est présentée par une seule partie. Ce type de tests prend la forme suivante :* $\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \leq \theta_0 \end{cases}$ *ou*

$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta > \theta_0 \end{cases}$

Et la région critique prend une des deux formes suivantes : $]-\infty, a]$ *ou* $[a, +\infty[$.

2. Le risque d'erreur

Un test statistique doit aboutir à choisir une des deux hypothèses H_0 et H_1. Il ya quatre solutions dont seulement deux sont justes :

1. H_0 est vraie et on a choisi H_0.
2. H_0 est fausse et on a rejeté H_0.
3. H_0 est vraie et on a rejeté H_0.
4. H_0 est fausse et on a choisi H_0.

On peut résumer ces différent cas de décisions dans le tableau suivant :

		Hypothèse vraie	Hypothèse vraie
		H_0	H_1
Hypothèse retenue	H_0	$1-\alpha$	β
Hypothèse retenue	H_1	α	$1-\beta$

Définition 2

On appelle risque de première espèce et on note α, la probabilité de rejeter à tort l'hypothèse nulle H_0 alors qu'elle est vraie.

$$\alpha = P(rejeter H_0 | H_0 vraie)$$

Le risque de première espèce α est aussi appelé " seuil de signification du test".

La quantité $1-\alpha$ est appelée niveau de confiance du test.

Définition 3

On appelle risque de deuxième espèce et on note β, la probabilité d'accepter l'hypothèse nulle H_0 alors qu'elle est fausse.

$$\beta = P(accepter H_0 | H_0 fausse)$$

β se détermine par un calcul de probabilité si H_1 est précisément définie (car α étant fixé).

La quantité $1-\beta$; qui est la probabilité de rejeter H_0 alors q'elle est fausse; est appelée "la puissance du test".

3. Le mécanisme général d'un test statistique

La méthodologie des tests consiste à l'aide des résultats expérimentaux à une question concernant les paramètres de la loi de probabilité des variables aléatoires.
La réalisation d'un test statistique passe par les étapes suivantes :

1. Position de La question biologique : On formule la problématique à l'aide d'une question simple de sorte qu'elle doit avoir que deux réponses possibles : oui ou non.

2. Formulation des hypothèses : une hypothèse nulle H_0 qui est toujours de non-effet ("il n'y a pas de différence entre ...", "il n'y a pas de relation entre ..."). Et une hypothèse alternative H_1, établie selon nos connaissances du domaine sous étude.
   • Si on ne connaît rien, H_1 est bilatérale : "il y a une relation entre...".
   • Si on a des connaissances plus détaillées, on peut parfois les utiliser dans le test, H_1 devient unilatérale: "il y a une relation positive entre...".

3. Choix du test : Il faut noter qu'il est nécessaire de définir tout d'abord le type de la variable étudiée, discrète ou continue, puis en fonction définir le nombre d'échantillons. Ensuite choisir parmis les différents tests disponibles, le test adéquat pour répondre à la problématique.

4. Calcul de la statistique du test : Après avoir défini le seuil de signification du test α (en général en prend $\alpha = 5\%$ ou $\alpha = 1\%$), on calcule la valeur cervée à partir de l'échantillon.

5. Prendre la décision : soit l'acceptation de l'hypothèse nulle H_0, soit le rejet de cette hypothèse.

6. Faire une interprétation des résultats

4. Les tests de conformité : Comparaison à une valeur théorique

Il s'agit de vérifier si les différences constatés entre la distribution théorique et la distribution expérimentale sont liées à la constitution de l'échantillon, via un paramètre donné. Ce paramètre soit toujours une des caractéristiques de la variable étudiée : sa moyenne, sa proportion ou sa variance.

4.1. Test de conformité d'une moyenne

On étudie une variable quantitative X et on cherche à savoir si les cervations (un échantillon de données de taille n, de moyenne cervée $\bar{X}$ et de variance σ_e^2) provenant d'une population de moyenne m et de variance σ^2 concordent avec une loi théorique de moyenne m_0.

Hypothèses à tester

$$\begin{cases} H_0 : m = m_0 \\ H_1 : m \neq m_0 \end{cases}$$

Statistique du test et règle de décision

1. variance σ^2 connue :

La statistique du test dans ce cas est : $Z = \dfrac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$

Sous H_0 : $z_c = \dfrac{\bar{X} - m_0}{\frac{\sigma}{\sqrt{n}}}$.

. Si $|z_c| > z_{1-\alpha/2}$: on rejette H_0 au seuil α.

. Si $|z_c| \leq z_{1-\alpha/2}$: on accepte H_0

2. variance σ^2 inconnue et $n \geq 30$:

$Z = f(S) \sim \mathcal{N}(0,1)$
Sous H_0 : $z = \dfrac{\bar{X} - m_0}{\frac{S}{\sqrt{n}}}$ où : $S^2 = \frac{n}{n-1}\sigma^2$ est la variance estimée.

. Si $|z_c| > z_{1-\alpha/2}$: on est dans la région critique et on doit rejeter H_0 au seuil α.

. Si $|z_c| \leq z_{1-\alpha/2}$: on accepte H_0

3. variance σ^2 inconnue et $n < 30$:

$T = \dfrac{\bar{X} - m}{\frac{S}{\sqrt{n}}} \sim \mathcal{T}(n-1)ddl.$
Sous H_0 : $t_c = \dfrac{\bar{X} - m_0}{\frac{S}{\sqrt{n}}}$.

. Si $|t_c| > t_{1-\alpha/2,n-1}$: on rejette H_0 au seuil α.

. Si $|t_c| \leq t_{1-\alpha/2,n-1}$: on accepte H_0

Le tableau suivant résume les différents cas d'un test de conformité d'une moyenne.

$H_0 : m = m_0 \;\; H_1 : m \neq m_0$			
$n \geq 30$		$n < 30$ et X normal	
σ connu	σ inconnu	σ connu	σ inconnu
Statistique $Z = f(\sigma) \sim \mathcal{N}(0,1)$ $z = \frac{\bar{X} - m_0}{\frac{\sigma}{\sqrt{n}}}$	Statistique $Z = f(S) \sim \mathcal{N}(0,1)$ $z = \frac{\bar{X} - m_0}{\frac{S}{\sqrt{n}}}$	Statistique $Z = f(\sigma) \sim \mathcal{N}(0,1)$ $z = \frac{\bar{X} - m_0}{\frac{\sigma}{\sqrt{n}}}$	Statistique $T = f(S) suit loi de Student$ $t = \frac{\bar{X} - m_0}{\frac{S}{\sqrt{n}}}$

4.2. Test de conformité d'une fréquence

On souhaite tester la conformité et la représentativité d'un échantillon de taille n, on étudie une variable X, qui représente le nombre k de succès parmi n tirages. On note $\hat{p} = \frac{k}{n}$ et on veut vérifier si les cervations concordent avec une loi théorique de probabilité de succès p_0. La loi théorique de X est une loi binomiale de probabilité p pour n tirages. On peut raisonner par un test de conformité où les hypothèses à tester sont : $\begin{cases} H_0 : p = p_0 \\ H_1 : p \neq p_0 \end{cases}$

Afin de pouvoir utiliser une loi de référence standard, la taille de l'échantillon doit être suffisamment grande ($n \geq 30$), dans ce cas le théorème de la limite centrale garantie la convergence de la loi binomiale vers la loi normale.

La statistique du test à calculer est alors :

$z_c = \dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ qui suit la loi normale centrée réduite $\mathcal{N}(0,1)$, et la valeur seuil est z_α.

Règle de décision

. Si $|z_c| > z_{1-\alpha/2}$: on est dans la région critique et on doit rejeter H_0 au seuil α.

. Si $|z_c| \leq z_{1-\alpha/2}$: on accepte H_0

4.3. Test de conformité d'une variance

On souhaite tester la conformité de la variance d'une population normale de variance σ^2 à partir d'un échantillon de taille n et de variance σ_e^2.
On définit une nouvelle variable aléatoire notée $\mathcal{X}^2$ appelée variable de "Khi deux" et définie par

$$\mathcal{X}^2 = (n-1)\frac{S^2}{\sigma^2}$$

où S^2 désigne l'estimateur de σ^2 à partir de l'échantillon. Cette variable aléatoire est une variable continue définie comme la somme des carrés de n variables aléatoire X_i telles que : $X_i \sim \mathcal{N}(0,1)$, $\forall i = \overline{1,n}$; son espérance mathématique est n et sa variance est $2n$.

In effectue le test suivant : $\begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 \neq \sigma_0^2 \end{cases}$

Sous H_0 la statistique du test est :

$$\mathcal{X}_c^2 = (n-1)\frac{S^2}{\sigma_0^2}$$

où la valeur seuil est $\mathcal{X}^2_{\alpha,n-1}$.

La règle de décision

• Pour un test bilatéral : $H_1 : \sigma^2 \neq \sigma_0^2$, la région d'acceptation de H_0 est donnée par l'intervalle

$$\left[\mathcal{X}^2_{\frac{\alpha}{2},n-1} \ ; \ \mathcal{X}^2_{1-\frac{\alpha}{2},n-1}\right]$$

• Pour un test unilatéral à gauche : $H_1 : \sigma^2 < \sigma_0^2$, la région d'acceptation est un intervalle de la forme :

$$\left]\mathcal{X}^2_{\alpha,n-1}\ ;\ +\infty\right[$$

c'est à dire on accepte H_0 si $\mathcal{X}^2_c > \mathcal{X}^2_{\alpha,n-1}$

• Pour un test unilatéral à droite : $H_1 : \sigma^2 > \sigma_0^2$, la région d'acceptation est un intervalle de la forme :

$$\left]-\infty\ ;\ \mathcal{X}^2_{1-\alpha,n-1}\right]$$

c'est à dire on accepte H_0 si $\mathcal{X}^2_c \leq \mathcal{X}^2_{1-\alpha,n-1}$

5. Les tests d'homogénéité : Comparaison de deux valeurs cervées

Ce type de tests s'intéresse à comparer directement deux valeurs expérimentales, au lieu de comparer une valeur cervée à une valeur d référence.

5.1. Homogénéité des moyennes

• **cas des grands échantillons** ($n \geq 30$)

On étudie deux variables X_1 et X_2 sur deux échantillons de tailles $n_1 (n_1 \geq 30)$ et $n_2 (n_2 \geq 30)$, de moyennes cervées $\bar{X}_1$, $\bar{X}_2$, et de variances cervées σ^2_{e1} et σ^2_{e2} issues respectivement de deux populations de moyennes m_1 et m_2 et de variances σ_1^2, σ_2^2.

On cherche à vérifier si ces cervations proviennent de la même loi théorique.

Les hypothèses à tester sont donc : $\begin{cases} H_0 : m_1 = m_2 \\ H_1 : m_1 \neq m_2 \end{cases}$

Statistique du test

On a $X_1 \sim \mathcal{N}(\bar{x_1}, \sigma^2_{e1})$ et $\bar{X_1} \sim \mathcal{N}(m_1, \frac{\sigma_1^2}{n_1})$

$X_2 \sim \mathcal{N}(\bar{x_2}, \sigma^2_{e2})$ et $\bar{X_2} \sim \mathcal{N}(m_2, \frac{\sigma_2^2}{n_2})$

La statistique du test est

$$Z = \frac{(\bar{X_1} - \bar{X_2}) - E(\bar{x_1} - \bar{X_2})}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$= \frac{(\bar{X_1} - \bar{X_2}) - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

On distingue deux cas :

1. Les variances σ_1^2 et σ_2^2 sont connues

Sous H_0 :

$$z_c = \frac{\bar{X_1} - \bar{X_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0, 1)$$

Règle de décision

. Si $|z_c| > z_{1-\alpha/2}$: on rejette H_0 et on conclut que les deux populations ont des moyennes différentes au seuil α.

. Si $|z_c| \leq z_{1-\alpha/2}$: on accepte H_0 et on constate que les moyennes des deux populations sont égales.

1. Les variances σ_1^2 et σ_2^2 sont inconnues

Dans ce cas on les remplace par leurs estimations respectivement :

$S_1^2 = \sigma^2_{e1}\frac{n_1}{n_1-1}$ et $S_2^2 = \sigma^2_{e2}\frac{n_2}{n_2-1}$

D'où la statistique du test est :

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim \mathcal{N}(0,1)$$

La règle de décision reste la même.

- **Cas des petits échantillons** ($n < 30$)

On sait que la distribution d'échantillonnage des moyennes d'échantillons de taille n, issus d'une population normale de moyenne m et d'écart-type σ, suit la loi normale $\mathcal{N}(m, \frac{\sigma}{n})$; ce résultat est vrai quelque soit n. Par suite, le rapport $\dfrac{\bar{X} - m}{\frac{\sigma}{n}}$ est une variable centrée réduite de Gauss. Si la population n'est pas normale, mais la taille de l'échantillon est suffisante ($n \geq 30$), le théorème de la limite centrale assure que la distribution des moyennes est approximativement gaussienne; Le rapport $\dfrac{\bar{X} - m}{\frac{\sigma}{n}}$ est lui aussi approximativement une variable centrée réduite de Gauss.

Cependant, en pratique il est rare que l'on connaisse la valeur de σ, on ne connait qu'une estimation S valeur calculée de l'estimateur S^2.

Que peut-on dire alors de la variable $\dfrac{\bar{X} - m}{\frac{\sigma}{n}}$?

Sous réserve que le caractère étudié soit distribué dans la population selon une loi normale, on peut démontrer que ce rapport suit une loi de Student à $(n-1)$ ddl, et que cette loi qui converge rapidement vers la loi de Gauss lorsque n augmente peut être remplacée par elle dès que $n \geq 30$. Pour le cas des petits échantillons et sous réserve que les échantillons proviennent des **populations normales** et de **mêmes variances** $\sigma_1^2 = \sigma_2^2$, la statistique du test est alors :

$$T = \frac{|\bar{X}_1 - \bar{X}_2|}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathcal{T}_\alpha(n_1 + n_2 - 2)ddl, \quad \hat{\sigma}^2 = \frac{n_1\sigma_{e1}^2 + n_2\sigma_{e2}^2}{n_1 + n_2 - 2}$$

la règle de décision :

On calcule $t_c = \dfrac{|\bar{X}_1 - \bar{X}_2|}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ et on la compare avec la valeur seuil de la loi de Student.

• Si $|t_c| > t_{1-\alpha/2,(n_1+n_2-2)}$: on rejette H_0 et on conclut que les deux populations ont des moyennes différentes au seuil α.

• Si $|t_c| \leq t_{1-\alpha/2,(n_1+n_2-2)}$: on accepte H_0 et on constate que les moyennes des deux populations sont égales.

5.2. Homogénéité des proportions

La démarche est la même que pour le cas précédent : on souhaite comparer deux population par rapport à la proportions d'individus pour lesquels la variable prend une certaine modalité "A". On tire de ces deux populations deux échantillons indépendants de taille respectives n_1 et n_2 sur lesquels On détermine les proportions d'individus de type "A". elles valent respectivement $\hat{p}_1 = \frac{k_1}{n_1}$ et $\hat{p}_2 = \frac{k_2}{n_2}$. On cherche à savoir si ces cervations proviennent de la même loi théorique, une loi binomiale de probabilité de succès p. Ici aussi, on utilise la convergence de la loi binomiale vers la loi normale.

On cherche à tester : $\begin{cases} H_0 : p_1 = p_2 = p \\ H_1 : p_1 \neq p_2 \end{cases}$

telle que p est estimée par $\hat{p} = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_1} = \frac{k_1+k_2}{n_1+n_2}$

Étant donné que $n_1 \geq 30$ et $n_2 \geq 30$, La statistique du test est alors :

$$z_c = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} \sim \mathcal{N}(0,1)$$

Règle de décision

On accepte H_0 si $|z_c| \leq z_{1-\alpha/2}$ fractile de la loi normale centrée réduite et alors la différence entre p_1 et p_2 n'est pas significative, sinon on la rejette et on constate que la différence entre p_1 et p_2 est significative au seuil α.

5.3. Homogénéité des variances : Test de Fisher

D'une Façon analogue aux deux précédente, on peut s'interroger sur l'égalité des variances σ_1^2 et σ_2^2 de deux populations indépendantes. $\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases}$

Pour cela on tire de ces deux populations deux échantillons indépendants de tailles respectives n_1 et n_1, de moyennes $\bar{X}_1$ e t$\bar{X}_2$ et de variances σ_{e1}^2 et σ_{e2}^2. Il n'est pas nécessaire que n_1 et n_1 soient grands, mais il est impératif que les deux populations soient normales.

on calcule la quantité

$$F_c = \frac{S_1^2}{S_2^2} \sim F_\alpha(n_1 - 1, n_2 - 1)$$

telle que : $S_1^2 \geq S_2^2 \quad , \quad S_1^2 = \frac{n_1}{n_1 - 1}\sigma_{e1}^2 \quad \text{et} \quad S_2^2 = \frac{n_2}{n_2 - 1}\sigma_{e2}^2$.

Règle de décision

Si $F_c \in [F_{\frac{\alpha}{2}}, n_1 - 1, n_2 - 1 \; F_{1-\frac{\alpha}{2}}, n_1 - 1, n_2 - 1]$, on accepte H_0 au risque α et on constates qu'il n ya pas une différence significative entre les variances des deux populations.

6. Analyse de la variance : ANOVA

Contrairement à ce que pourrait laisser penser son nom, l'analyse de la variance n'est pas une méthode qui permet d'étudier les différences de variances entre populations, mais une méthode pour étudier les différences de moyenne entre populations (par exemple, trois populations ont-elles la même moyenne? ou autrement dit, les différences de moyenne entre les trois populations sont-elles significatives ?). Cette méthode, néanmoins, doit son nom au fait qu'elle utilise des mesures de variance afin de déterminer le caractère significatif, ou non, des différences de moyenne mesurées sur les populations.
Il s'agit d'une généralisation à k populations du classique test de comparaison de moyennes de deux ééchantillons : le célèbre test de T.

- **Idée générale**

L'idée de l'analyse de la variance repose sur un modèle qu'on se donne a priori des données. On suppose ainsi, par exemple, qu'une variable mesurée $\mathbf{Y}$ vérifie une relation linéaire avec un ensemble de p variables explicatives dénotées $\mathbf{X}_i$. La relation est du type suivant :

$$Y = \mu + \sum_{i=1}^{p} \alpha_i X_i + \varepsilon$$

avec:

- μ un paramètre commun à toutes les cervations, c'est-à-dire une ordonnée à l'origine (dont on pourra tester éventuellement la nullité plus tard).

- ε représente la variabilité aléatoire du modèle, non contrôlable.

On s'attache ensuite à l'étude de la contribution de ces différents termes à la variance de $\mathbf{Y}$, grâce à une décomposition dite de **l'analyse de la variance**.

Conditions d'application de l'ANOVA

. Le caractère étudié suit la loi normale.

. Les variances des populations sont toutes égales (**Homoscédasticité**).

. Les échantillons sont prélevés aléatoirement et indépendamment dans les populations.

Noter Bien

Il est important de comprendre que l'ANOVA n'est pas un test pour classer des moyennes, mais plutôt c'est un test permettant de comparer les moyennes de différents groupes et dire, si parmi l'ensemble, au moins une d'entre elles diffère des autres, mais on ne sait ni laquelle ni combien d'entre elles.

Analyse de variance à un facteur contrôlé

On parle d'**ANOVA** à un facteur lorsque les groupes analysés se distinguent par un seul facteur qualitatif (par exemple, Biologiste et statisticien). Elle a pour but de comparer les moyennes de n populations à partir d'échantillons aléatoires et indépendants prélevés dans chacune d'elles.

Hypothèses à tester

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

Procédure des calculs

Il s'agit en pratique de décomposer la variabilité selon (au moins) deux critères :

Variabilité non expliquée, ou résiduelle, entre un terme estimé et la vraie valeur mesurée, qu'on appellera **SCR**, on parle aussi de variance intra-classe. Variabilité expliquée par le modèle, c'est-à-dire la différence entre l'estimation de moyenne d'une classe et la moyenne totale des cervations, qu'on appellera **SCF**, pour la variance due au facteur **A**, c'est la variance inter-classe.
Cette décomposition se fait par **l'équation d'analyse de variance suivante** :

$$\underbrace{\sum_{I=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2}_{variationtotale} = \underbrace{\sum_{I=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\bar{y_i})^2}_{variationINTRA} + \underbrace{\sum_{I=1}^{I} n_i(\bar{y_i}-\bar{y})^2}_{variationINTER}$$

$$\mathbf{SCT} = \mathbf{SCR} + \mathbf{SCF}$$

tels que :

- n : nombre total d'individus.
- y_{ij} : cervation de la variable endogène Y.
- n_i : nombre d'individus du niveau i.
- $\bar{y_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$: la moyenne des individus du niveau i.
- $\bar{y} = \frac{1}{n}\sum_{I=1}^{I}\sum_{j=1}^{n_i} y_{ij} = \frac{1}{n}\sum_{I=1}^{I} n_i\bar{y_i}$: La moyenne totale.

Si le facteur A a un effet su la variable endogène Y, la variation INTER sera importante par rapport à la variation INTRA.
À partir de cette définition, on va comparer les espérances des variances **SCF** et **SCR** en faisant leur rapport. Il se trouve (comme on peut le voir dans la décomposition mathématique) que les deux termes sont tous les deux une estimation de la variabilité résiduelle si le facteur A n'a pas d'effet. De plus, ces deux termes suivent chacun une loi de **khi-deux**, **leur rapport suit donc une loi de F** (voir plus loin pour les degrés de liberté de ces lois).

Tableau d'ANOVA

Source de variation	Somme des carrés SC	ddl	Moyenne somme des carrés (MC)	Statistique F de Fisher
INTER(due au facteur)	$SCF = \sum_{I=1}^{I} n_i(\bar{y}_i - \bar{y})^2$	$I-1$	$MCF = \frac{SCF}{I-1}$	$F_c = \frac{MCF}{MCR}$
INTRA (résiduelle)	$SCR = \sum_{I=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$	$n-I$	$MCR = \frac{SCR}{n-I}$	
Totale	$SCT = \sum_{I=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$	$n-1$	$MCT = \frac{SCT}{n-1}$	

La règle de décision

Au risque α, l'hypothèse H_0 sera rejetée si $F_c > F_t = F_{I-1,n-I,\alpha}$ où $F_{I-1,n-I,\alpha}$ est la valeur seuil de la loi de Fisher.