Chapitre 04 : Analyse de la variance

I-ANALYSE DE LA VARIANCE A UN CRITERE DE CLASSIFICATION

- **Les aspects descriptifs**

- **Principes généraux**

Les notions de **modèle observé** et de **tableau d'analyse de la variance (ANOVA)**, et la réalisation pratique de l'analyse. Il nous paraît en effet important, tant pour la compréhension que pour l'utilisation de l'analyse de la variance, de conserver une certaine aptitude au calcul «manuel », en particulier en ce qui concerne la détermination des **sommes des carres des écarts (SCE)**.

La décomposition de la variation totale

1◦ Nous supposerons qu'on dispose au départ de **p échantillons** ou **séries d'observations**, **d'effectifs ni (i = 1, . . ., p)**, et nous désignerons l'effectif total par **n.** :

$$\boxed{n_. = \sum_{i=1}^{p} n_i \,.}$$

2◦ Nous désignerons aussi les différentes observations par le symbole x_{ik} (*i* = 1 , . .,*p)* et (*k*

x_{ik} $k^{\text{ème}}$ = 1 , . . . , *ni*), la valeur étant donc la observation du i^{eme} ´echantillon. On peut

3◦ En fonction de ces différents éléments, il est possible de subdiviser les écarts entre les observations individuelles et la moyenne générale en deux composantes additives :

$$\boxed{x_{ik} - \bar{x} = (\bar{x}_i - \bar{x}) + (x_{ik} - \bar{x}_i)}.$$

La composante globale est appelée **variation totale** et les deux composantes partielles sont appelées, d'une part, **variation factorielle** ou liée au **facteur contrôlé**, ou encore entre échantillons, et d'autre part, **variation résiduelle** ou dans les échantillons.

4◦ En élevant au **carré** les deux membres de la relation précédente, et en **sommant** pour toute les valeurs observées, on obtient **l´´equation d'analyse de la variance**:

$$\sum_{i=1}^{p}\sum_{k=1}^{n_i}(x_{ik} - \bar{x})^2 = \sum_{i=1}^{p}\sum_{k=1}^{n_i}(\bar{x}_i - \bar{x})^2 + \sum_{i=1}^{p}\sum_{k=1}^{n_i}(x_{ik} - \bar{x}_i)^2$$

ou $$\boxed{\sum_{i=1}^{p}\sum_{k=1}^{n_i}(x_{ik} - \bar{x})^2 = \sum_{i=1}^{p}\left[n_i(\bar{x}_i - \bar{x})^2\right] + \sum_{i=1}^{p}\sum_{k=1}^{n_i}(x_{ik} - \bar{x}_i)^2}.$$

On constate ainsi que la somme des carrés des écarts par rapport à la moyenne générale, est appelée **somme des carres des écarts totale : SCEt**, peut elle aussi être divisée en deux composantes additives : **une somme des carrés des écarts factorielle : SCEa** ou entre échantillons, et une **somme des carrées des écarts résiduelle SCEr**, on peut résumer l´´equation d'analyse de la variance sous la forme :

$$\boxed{\mathrm{SCE}_t = \mathrm{SCE}_a + \mathrm{SCE}_r}.$$

5◦ Des nombres de **degrés de liberté** : **ddl** peuvent être associés aux différentes sommes des carrés des écarts. Ces nombres de degrés de liberté sont aussi **additifs** et se présentent de la manière suivante :

$$\boxed{n_{.} - 1 = (p-1) + (n_{.} - p)}.$$

6◦ En divisant les **sommes des carrés des écarts** par leurs **nombres de degrés de liberté** respectifs, on définit des quantités appelées **carré moyen total :**

CMt, un **carré moyen factoriel : CMa** ou entre échantillons, et un **carré moyen résiduel : CMr** ou dans les échantillons:

$$\boxed{\mathrm{CM}_t = \mathrm{SCE}_t/(n_. - 1)}, \quad \boxed{\mathrm{CM}_a = \mathrm{SCE}_a/(p - 1)} \quad \text{et} \quad \boxed{\mathrm{CM}_r = \mathrm{SCE}_r/(n_. - p)}.$$

Ces carrés moyens sont aussi appelés **variances** et ils possèdent d'ailleurs certaines des propriétés des variances, notamment en ce qui concerne **leurs distributions d'´echantillonnage**.

7◦ Tableau d'analyse de la variance (ANOVA) : un critère de classification ou à un seul Facteur

Sources de variation	Degrés de liberté	Sommes des carrés des écarts	Carrés moyens
Différences entre échantillons	$p-1$	SCE_a	CM_a
Différences entre observations (dans les échantillons)	$n_. - p$	SCE_r	CM_r
Totaux	$n_. - 1$	SCE_t	

8◦ Le rapport des sommes des carrés des écarts factorielle sur la somme carrés des écarts totale permet de définir facilement le rapport de corrélation, aussi appelé **coefficient de corrélation non linéaire** :

$$\boxed{\eta = \sqrt{\mathrm{SCE}_a/\mathrm{SCE}_t}}.$$

x_i D'une manière générale, ce paramètre joue, dans le cas d'une **relation** liant les **différents échantillons** et les **différentes observations**. Le rapport de **corrélation** est toujours compris

entre **0 et 1**. Il est égal à **0** quand toutes les **moyennes sont ´égales entre elles**, et il est égal à **1** quand les **variances des différents échantillons sont toutes nulles**.

Certains logiciels associent systématiquement le carré du rapport de corrélation à toutes les analyses de la variance, en utilisant la notation r^2 ou R^2, et non pas η^2 ou tout autre symbole particulier. S'il s'agit bien l`a d'un paramètre jouant un rôle comparable `a celui du **coefficient de détermination**, il y a lieu toutefois d'être attentif au fait qu'il ne s'agit nullement, d'une façon générale, du carré d'un **coefficient de corrélation classique**.

Application

Dans cet exemple, nous allons vérifier s'il existe ou non, en moyenne, des différences significatives de hauteurs entre les trois types de forêts, et chiffrer éventuellement ces différences. Les hauteurs en mètre de 37 arbres sont mentionnées dans le tableau 1 suivant :

Hauteurs			Rangs		
Type 1	Type 2	Type 3	Type 1	Type 2	Type 3
23,4	22,5	18,9	8	5,5	1
24,4	22,9	21,1	12,5	7	2
24,6	23,7	21,2	16,5	10	3
24,9	24,0	22,1	18	11	4
25,0	24,4	22,5	19	12,5	5,5
26,2	24,5	23,6	23	14,5	9
26,3	25,3	24,5	25	20	14,5
26,8	26,0	24,6	29,5	21	16,5
26,8	26,2	26,2	29,5	23	23
26,9	26,4	26,7	31,5	26	27,5
27,0	26,7	–	33	27,5	–
27,6	26,9	–	35	31,5	–
27,7	27,4	–	36	34	–
–	28,5	–	–	37	–
Totaux			316,5	280,5	106

Tableau1. Comparaison des hauteurs des arbres de trois types de hêtraies : hauteurs Observées, en mètres et rangs.

Les moyennes correspondantes sont :

$\bar{x}_1$ = 25,97 (23,4+24,4+…….+27,7)/13

$\bar{x}_2$ = 25,39 (22,5+22,9+.......+28,5)/14 ,
$\bar{x}_3$ = 23,14 (18,9+21,1+.......+26,7)/10 et la moyenne générale
$\bar{x}$ = 24,98 m. (25,97x 13+(25,39x14)+(23,14x10)/37

Prenant la première observation du premier échantillon (**x_{11} = 23,4**), le modèle observé d'analyse de la variance s´´ecrit :

$$\boxed{x_{ik} - \bar{x} = (\bar{x}_i - \bar{x}) + (x_{ik} - \bar{x}_i)}.$$

SCE*t* =165,5198

L'écart négatif de **1,58 m** entre cette observation particulière et la moyenne générale provient, à la fois, du fait que l'endroit considère appartient à un type de forêts dont la moyenne est supérieure de **0,99 m**, par rapport à la moyenne générale, et que cet endroit présente une hauteur inferieure de **2,57 m**, par rapport à la moyenne de toutes les observations relatives à ce type de forêts.
Un calcul similaire pourrait être réalise pour chacune des 36 autres valeurs.

En sommant les carrés des écarts ainsi obtenus, on aboutirait aux trois sommes des carrés des écarts définies précédemment : pour cela on obtient les sommes des carrés de la manière suivante :

SCEt = $(-1,58)^2 + (-1,58)^2 + (-0,58)^2 + (-0,38)^2 + \ldots\ldots + (1,82)^2 = 165,53$
SCEa = $(0,99)^2 * 13 + (0,41)^2 * 14 + (1,84)^2 * 10 = 48,88$
SCEr = $(-2,57)^2 + (-1,57)^2 + (-1,37)^2 + (-1,07)^2 + \ldots\ldots + (+3,56)^2 = 116,53$

Cette façon de procéder est pour bien saisir le mécanisme de l'analyse de la variance.

Le tableau ci-dessous présente les sommes des carrés des écarts qui sont ainsi obtenues, les nombres de degrés de liberté et les carrés moyens.

$$\boxed{\mathrm{CM}_t = \mathrm{SCE}_t/(n_. - 1)}, \quad \boxed{\mathrm{CM}_a = \mathrm{SCE}_a/(p-1)} \quad \text{et} \quad \boxed{\mathrm{CM}_r = \mathrm{SCE}_r/(n_. - p)}.$$

On applique ces 3 formules on obtient les carrés moyens respectifs dont **n=37 arbres et p=3**

Sources de variation	Degrés de liberté	Sommes des carrés des écarts	Carrés moyens
Différences entre types de hêtraies	2	48,88	24,44
Différences entre observations (dans les types de hêtraies)	34	116,65	3,431
Totaux	36	165,53	

- **Tableau d'analyse de la variance de la comparaison des hauteurs moyennes desarbres de trois types de hêtraies**

Ce carré moyen total (ou cette **variance totale**), auquel correspond un écart type égal `a **2,14 m** (c'est la **racine carrée du CMt**), mesure globalement **l'hétérogénéité des hauteurs**, sans tenir compte de la subdivision en trois types de forêts. Le carré moyen résiduel (ou la variance résiduelle), auquel est associé un écart-type égal à **1,85 m** (**racine carrée du CMr**), mesure, toujours globalement, **l'hétérogénéité des hauteurs** à l'intérieure des trois types de forêts. Il faut rappeler que

Par définition, l'écart-type est la moyenne quadratique des écarts à la moyenne $\bar{x}$. On le note habituellement s (de l'anglais *standard deviation*) :

$$\sqrt{\frac{1}{n}\sum_i (x_i - \bar{x})^2} \qquad \{1\}$$

En outre, on peut compléter l'analyse par le calcul du **rapport de corrélation** ou de son carre :

$$\boxed{\eta = \sqrt{\mathrm{SCE}_a / \mathrm{SCE}_t}} .$$

$$\eta^2 = 48{,}88/165{,}53 = 0{,}30.$$

Ce paramètre mesure le **degré de dépendance** de la variable quantitative **« hauteur des arbres »** en fonction de la caractéristique nominale **« type de hêtraies »**. Comme un coefficient de détermination, il indique que **30 % de la variation totale** peut être expliquée par **les différences entre types de forets.**

La réalisation de l'analyse de la variance

1◦ *N*ous donnons à titre indicatif, quelques informations relatives à la **réalisation de l'analyse de la variance** en. Cette réalisation consiste essentiellement en une suite de déterminations de **sommes de carrés**

d'écarts (SCE), semblables à celle qui peut être effectuée pour toute **série d'observations**

2◦ En ce qui concerne la **somme des carrés des écarts résiduelle**, on peut calculer **séparément** les sommes des carrés des écarts relatives aux différents échantillons ou séries d'observations, et **somme**r ensuite les résultats ainsi obtenus. Si on désigne par ***Xi.*** et **SCE*i*** , respectivement, les sommes et les sommes des carrés des écarts relatives aux différentes séries d'observations, on a :

$$X_{i.} = \sum_{k=1}^{n_i} x_{ik} \quad \text{et} \quad \mathrm{SCE}_i = \sum_{k=1}^{n_i} x_{ik}^2 - X_{i.}^2/n_i \quad (\text{pour tout } i)\,,$$

ainsi que :

$$\mathrm{SCE}_r = \sum_{i=1}^{p} \mathrm{SCE}_i\,.$$

En réalité, la détermination des **sommes des carrés des écarts individuelles SCE*i*** n'est pas indispensable, en vue de calculer la **somme des carrés des écarts résiduelle**, mais cette détermination permet d'obtenir facilement les **variances** des différentes **séries d'observations** et donc de **comparer ces variances**, préalablement à toute inférence statistique.

3◦ Quant à la ***somme des carrés des écarts totale*** **(SCE*t*)**, on a, toujours par analogie avec le cas d'une **seule série d'observations** :

$$\mathrm{SCE}_t = \sum_{i=1}^{p}\sum_{k=1}^{n_i} x_{ik}^2 - X_{..}^2/n_{.}\,,$$

Le symbole ***X..,*** désignant la **somme de l'ensemble des *n.* observations** :

$$X_{..} = \sum_{i=1}^{p}\sum_{k=1}^{n_i} x_{ik} = \sum_{i=1}^{p} X_{i.}\,.$$

On remarquera que le premier terme qui intervient dans la relation relative à la somme des carrés des écarts totale **SCE*t*** n'est autre que la somme des premiers termes qui se présentent dans l'expression relative aux sommes des carrés des écarts individuelles **SCE*i***.

4◦ Enfin, la ***somme des carrés des écarts factorielle*** (**SCEa**) peut être obtenue soit par différence :

Soit par la relation :

$$\boxed{\mathrm{SCE}_a = \sum_{i=1}^{p} (X_{i.}^2/n_i) - X_{..}^2/n_. \, .}$$

Application : Dans l'exemple du tableau 1 comparaison des hauteurs des arbres de 3 types de hêtraies : **réalisation de l'analyse de la variance**.

On calcul les SCE individuelles séparément selon la formule :

$$\mathrm{SCE}_i = \sum_{k=1}^{n_i} x_{ik}^2 - X_{i.}^2/n_i \quad \text{(pour tout } i\text{)}$$

SCE1 = (23,4²+24,4²+……….+27,7²) - (23,4+24,4+………+27,7)²= 8.789,36 − 337,62²/13 = **22,15** **SCE2** = (22,5²+22,9²+……….+29,5²) - (22,5+22,9+… ……………………………………………………………………… +29,5)²=9.062,96 − 355,42²/14 = **40,88**

SCE3 = (18,9²+21,1²+……….+26,7²) - (18,9+21,1+……….+26,7)²= 5.408,22 −231,42²/10 = **53,62**

$$\mathrm{SCE}_r = \sum_{i=1}^{p} \mathrm{SCE}_i \, .$$

SCE*r* = **22,15** + **40,88** + **53,62** = **116,65** ,

$$\sum_{i=1}^{p} \sum_{k=1}^{n_i} (x_{ik} - \bar{x})^2$$

SCEt=(23,4−24,98)²+(24,4−24,98)²+(24,6−24,98)²+… ……………………………………………………………………………… +(27,7−24,98)²+(22,5 −24,98)² +(22,9−24,98)²+(23,7−24,98)²+…………..…………….+(28,5-24,98)²+(18,9−24,98)²+(21,1−24,98)² + (21,2− 24,98)²+…………..+ (26,7− 24,98)²= (23.260,54 − 924,42²/37 = **165,53** **SCEt**=**165,53** et

$$\boxed{F_{obs} = \mathrm{CM}_a/\mathrm{CM}_r}.$$

5° Le *test* de l'hypothèse nulle émise nécessite le calcul de la quantité :

Le rejet de l'hypothèse, au niveau de **probabilité *α*** , intervient quand cette quantité est trop élevée, c'est-à-dire quand :

$$P(F \geq F_{obs}) \leq \alpha \quad \text{ou} \quad F_{obs} \geq F_{1-\alpha} \,,$$

avec ***p* − 1** et ***n.* − *p* degrés de liberté**. Le caractère unilatéral du test résulte de ce que, dans tous les cas ou l'hypothèse nulle est fausse, les valeurs ***Fobs*** dépassent en moyenne les valeurs que donnent normalement les distributions ***F* de Fisher-Snedecor**.

Le **rejet de l'hypothèse nulle**, relative à un ensemble de *p* moyennes, soulève la question de savoir quelles sont les moyennes qui différent significativement les unes des autres

En outre, on peut calculer comme suit des limites de confiance, pour les moyennes $\boldsymbol{m_i}$ et pour les différences de moyennes *mi –mi'* :

$$\boxed{\bar{x}_i \pm t_{1-\alpha/2} \sqrt{\mathrm{CM}_r/n_i}} \quad \text{et} \quad \boxed{\bar{x}_i - \bar{x}_{i'} \pm t_{1-\alpha/2} \sqrt{\mathrm{CM}_r\,(1/n_i + 1/n_{i'})}} \,,$$

La variable ***t* de Student** étant une variable à ***n* − *p* degrés de liberté**. Ces formules sont semblables à celles qui concernent une ou deux populations, la seule différence étant que les estimations antérieures de la variance $\boldsymbol{\sigma^2}$ sont remplacées ici par $\mathbf{CM}_r$.

Des **limites de confiance** relatives à la **variance $\boldsymbol{\sigma^2}$** et à **l'écart-type $\boldsymbol{\sigma}$** peuvent également être obtenues selon les procédures habituelles, à partir de la **somme des carrés des écarts** ou du **carré moyen résiduel**, et grâce à la distribution $\boldsymbol{\chi^2}$ à ***n* – *p* degrés de liberté**.

Application : l'exemple du tableau1 : comparaison des hauteurs des arbres de 3 types hêtraies suite de l'analyse de la variance :
Nous pouvons maintenant clôturer l'analyse de la variance que nous avons entamée ` A partir du tableau 1, on obtient selon la formule de la **Fobs :**

$$\boxed{F_{obs} = \mathrm{CM}_a/\mathrm{CM}_r} \,.$$

$$\boxed{\mathrm{CM}_t = \mathrm{SCE}_t/(n_. - 1)} \,, \quad \boxed{\mathrm{CM}_a = \mathrm{SCE}_a/(p - 1)} \quad \text{et} \quad \boxed{\mathrm{CM}_r = \mathrm{SCE}_r/(n_. - p)} \,.$$

CM_a= carré moyen factoriel= **48,88**/3-1=24,44 CM_r**= carré moyen résiduel=116,65/37-**

3=3,43 **F**$_{obs}$=24,44/**3,43=7,12** et
P(F ≥ 7,12) = 0,0026

Les limites de confiance des différences sont, pour un degré de confiance égal à 0,95 et pour les deux premiers types de forets :

$$\boxed{\bar{x}_i - \bar{x}_{i'} \pm t_{1-\alpha/2}\sqrt{\mathrm{CM}_r\,(1/n_i + 1/n_{i'})}}\,,$$

$$25{,}97 - 25{,}39 \pm 2{,}032\,\sqrt{3{,}431\,(1/13 + 1/14)} = 0{,}58 \pm 1{,}45 = -0{,}87 \text{ et } 2{,}03 \text{ m}\,,$$

pour le premier et le troisième type de forêts :

$$25{,}97 - 23{,}14 \pm 2{,}032\,\sqrt{3{,}431\,(1/13 + 1/10)} = 2{,}83 \pm 1{,}58 = 1{,}25 \text{ et } 4{,}41 \text{ m}\,,$$

$$25{,}39 - 23{,}14 \pm 2{,}032\,\sqrt{3{,}431\,(1/14 + 1/10)} = 2{,}25 \pm 1{,}56 = 0{,}69 \text{ et } 3{,}81 \text{ m}\,.$$ et pour les deux derniers types de forêts :

Le fait que le premier intervalle de confiance englobe la valeur **zéro** indique qu'il n'y a pas de différence significative entre les deux premiers types de hêtraies, ce qui était déjà la

$$\bar{x}_{12} = (337{,}6 + 355{,}4)/27 \;\text{ ou }\; \big[13\,(25{,}97) + 14\,(25{,}39)\big]/27 = 25{,}67 \text{ m}\,,$$ con

clusion de l'exemple. On peut en conséquence calculer éventuellement une moyenne globale pour l'ensemble de ces deux types :

et déterminer des limites de confiance relatives à la différence entre cette nouvelle moyenne et la moyenne du troisième type de forets :

En vue de tenir compte du fait qu'on procède en réalité `a trois comparaisons, dans la détermination des trois intervalles de confiance initiaux, on aurait pu remplacer la valeur ***t* classique** $(t_{0,975} = 2{,}032)$, par une valeur ***t*** définie au sens de **Bonferroni**: $t_{1-0,05/6} = t_{0,99167} = 2{,}518$.
Cette façon de faire aurait conduit à étendre assez sensiblement les différents intervalles de confiance, sans modifier, dans le cas présent, les conclusions finales.

-RESUME DE L'ANOVA A UN CRITERE DE CLASSIFICATION

• ANOVA à un facteur - Introduction

- **Analyse de la variance :**

L'analyse de la variance a pour but 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

: Les variances sont homogènes

: Au moins une des variances est différente des autres

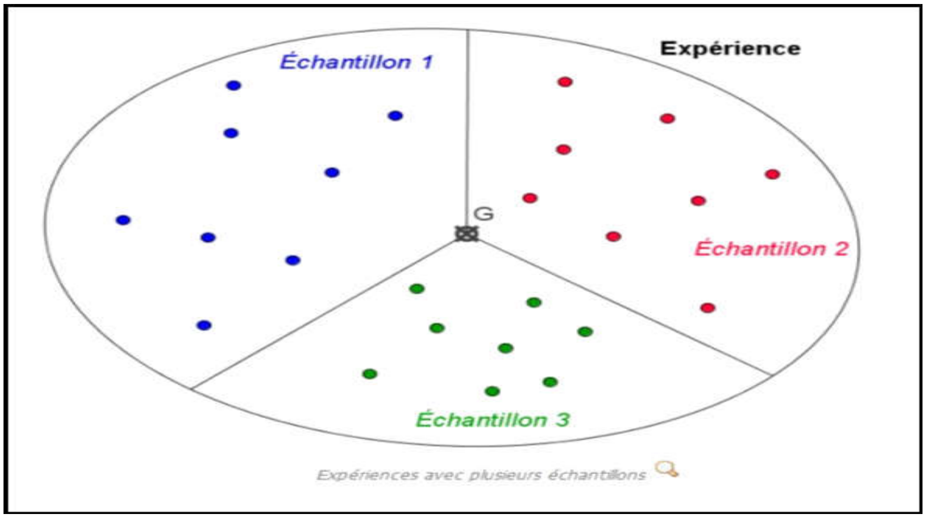
✓ Utilisation d'un test de comparaison de plusieurs variances

- **Conclusion**

$(H_0)_\sigma$: Est rejetée : il est théoriquement **impossible** de **comparer des échantillons** qui ne varient pas de la même manière.

$(H_1)_\sigma$ **:** N'est **pas rejetée** : par conséquent, il est **possible de comparer les moyennes de tels échantillons**

- **Expérience avec k échantillons - Données initiales**

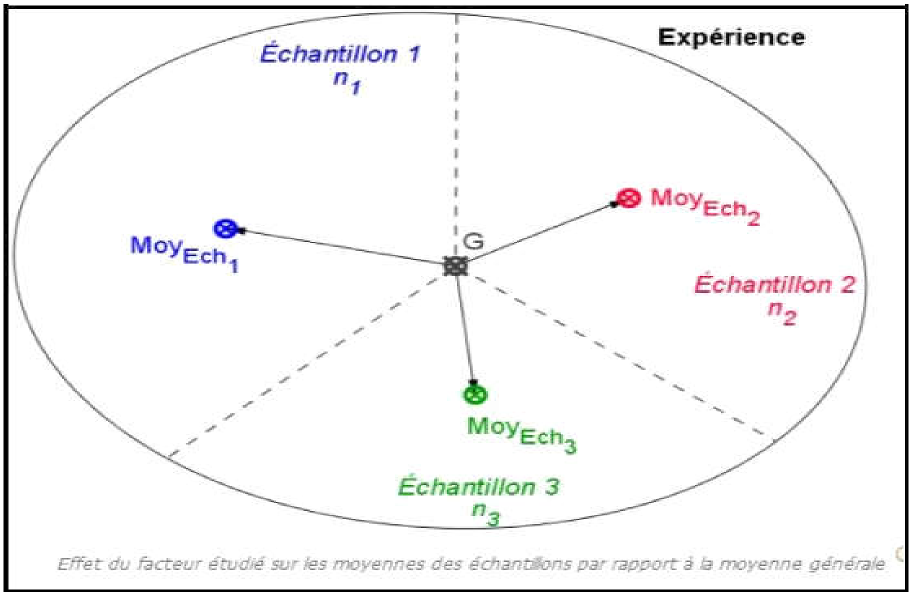


- Soit une Expérience faisant intervenir ***k*** échantillons de n_i individus.
- Le nombre total d'individus est $N = \Sigma n_i$
- On calcule la moyenne générale des mesures de l'expérience **(G)**.

Variabilité totale

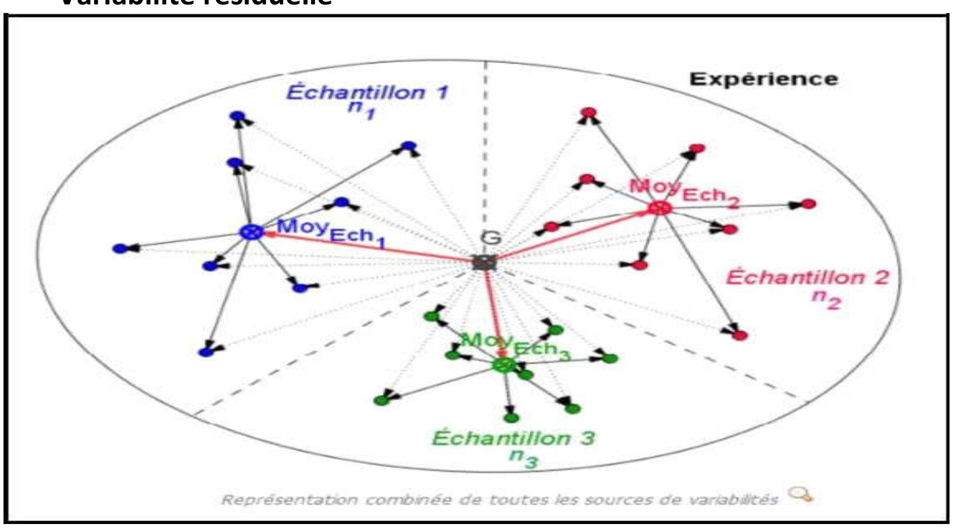
- Variabilité totale au sein de l'expérience (quel que soit l'échantillon) : reflète les écarts de tous les individus par rapport à la moyenne générale (G) de l'expérience.
- Calcul de la Somme des Carrés des Écarts à la moyenne totale SCE_r
- Degrés de liberté **(DDL)** associés : ***N-1.***

Variabilité factorielle



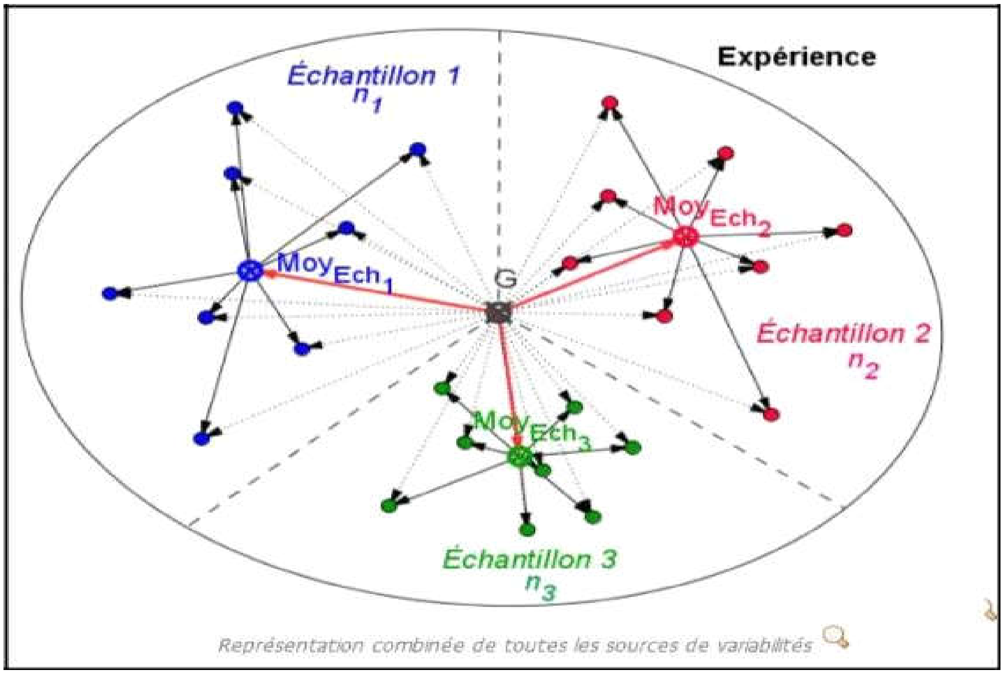
- Variabilité factorielle : reflète les écarts des moyennes des échantillons (supposées influencées par le facteur étudié) par rapport à la moyenne générale (G) de l'expérience.
- Calcul de la Somme des Carrés des Écarts à la moyenne factorielle ***(SCE$_F$)***
- DDL associés : **k-1**

Variabilité résiduelle



- Variabilité résiduelle (liée à l'individu) : reflète l'importance des variations individuelles dans chaque échantillon.
- Calcul de la Somme des Carrés des Écarts à la moyenne résiduelle ***SCE$_R$***
- DDL associés : ***N-k.***

Bilan



Représentation combinée de toutes les sources de variabilités

Pour résumer :

- $SCE_T = SCE_F + SCE_R$
- DDL associés : $N-1 = k-1 + N-k$.
- On comparera les variabilités factorielles $s^2_F = SCE_F/k-1$ *et résiduelle* $s^2_R = SCE_R/N-K$

Comparaison des moyennes - Hypothèses

H0 : toutes les moyennes sont identiques

H1 : au moins une des moyennes est différente des autres

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

Pour l'ensemble de l'expérience :

- Taille totale

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

- Moyenne générale

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

- Variance totale

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{\bar{x}}\right)^2$$

Variance factorielle (variance intergroupe) :

- Dispersion des valeurs d'un échantillon à l'autre (influence du facteur) :

$$s_F^2 = \frac{1}{k-1} \sum_{i=1}^{k} n_i \cdot \left(\bar{x}_i - \bar{\bar{x}}\right)^2$$

Variance résiduelle (variance intragroupe)

- Dispersion des valeurs à l'intérieur des échantillons (variabilité individuelle) :

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

ANOVA à un facteur - Conclusion

Tableau 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	$CM_T = s^2$	

$SCE_T = SCE_F + SCE_R$

- F suit une loi de Snédécor à $\nu_1 = k-1$ et $\nu_2 = n-k$ 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 H_0 à tort)

Décision

$F > f_\alpha$ H_0 **Si** => rejet de au risque α :

La **variance factorielle** est significativement **supérieure à la variance résiduelle** : les moyennes diffèrent significativement entre-elles. → on attribue une influence **significative au facteur étudié**.

Recherche du degré de signification p (recherche du risque α 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.**

II-ANALYSE DE LA VARIANCE A DEUX CRITERES DE CLASSIFICATION

L'analyse de la variance à deux critères de classification

- Introduction

1◦ L'analyse de la variance à deux critères de classification

Les deux facteurs envisagés peuvent être soit placés sur **pied d'égalité** dans ce cas les modèles d'analyse de la variance sont dits ***croisés***, soit au contraire **subordonnés l'un à l'autre** les modèles sont dits ***hiérarchisés***. Le cas hiérarchique est parfois qualifié aussi de ***multi-niveaux.***

.

Dans les différents cas, on doit également faire la distinction entre les modèles ***fixes***, les modèles
aléatoires et les modèles ***mixtes*.**

2◦ *N*ous considérerons tout d'abord les ***aspects descriptifs***, puis les ***aspects inferentiels*** de l'analyse à deux critères, en nous limitant dans un premier temps aux modèles **croisés à effectifs égaux**. Nous envisagerons ensuite les ***modèles croisés à effectifs inégaux*** et les ***modèles hiérarchisés***.

3◦ Les conditions d'application sont: ***populations normales*** *et de* ***même variance****,* et ***échantillons aleatoires****, simples et indépendants*.

- **Les modèles croisés à effectifs égaux : Aspects descriptifs**

- **La décomposition de la variation totale**

Considérons ***p q*** **échantillons** ou **séries d'observations** de **même effectif** ***n***, et désignons les **observations individuelles** par x_{ijk}, les indices *i*, *j* et *k* étant relatifs respectivement aux **différentes modalités du premier critère de classification** ($i = 1, \ldots, p$), aux différentes **modalités du deuxième critère de classification** ($j = 1, \ldots, q$), et aux **différentes observations d'un même échantillon** ou d'une même série ($k = 1, \ldots, n$).

$$\boxed{\bar{x}_{ij.} = \frac{1}{n}\sum_{k=1}^{n} x_{ijk}},$$

A partir de telles données, on peut **calculer différentes moyennes**, à savoir une moyenne pour chacun des échantillons ou séries d'observations ($i = 1, \ldots, p$ et $j = 1, \ldots, q$) :

Une moyenne pour chacune des modalités de chacun des deux critères de classification (*i* = 1 , . . . , *p*

$$\boxed{\bar{x}_{i..} = \frac{1}{q\,n}\sum_{j=1}^{q}\sum_{k=1}^{n} x_{ijk} = \frac{1}{q}\sum_{j=1}^{q}\bar{x}_{ij.}} \quad \text{et} \quad \boxed{\bar{x}_{.j.} = \frac{1}{p\,n}\sum_{i=1}^{p}\sum_{k=1}^{n} x_{ijk} = \frac{1}{p}\sum_{i=1}^{p}\bar{x}_{ij.}},$$

d'une part, et *j* = 1 , . . . , *q* d'autre part) :

et une moyenne générale :

$$\boxed{\bar{x}_{...} = \frac{1}{p\,q\,n}\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n} x_{ijk} = \frac{1}{p\,q}\sum_{i=1}^{p}\sum_{j=1}^{q}\bar{x}_{ij.} = \frac{1}{p}\sum_{i=1}^{p}\bar{x}_{i..} = \frac{1}{q}\sum_{j=1}^{q}\bar{x}_{.j.}}.$$

$$\boxed{\begin{aligned} x_{ijk} - \bar{x}_{...} &= (\bar{x}_{ij.} - \bar{x}_{...}) + (x_{ijk} - \bar{x}_{ij.}) \\ &= (\bar{x}_{i..} - \bar{x}_{...}) + (\bar{x}_{.j.} - \bar{x}_{...}) + (\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x}_{...}) + (x_{ijk} - \bar{x}_{ij.}) \end{aligned}}.$$

- **subdivision des écarts par rapport à la moyenne générale :** en **deux**, puis en **quatre composantes :**

La première décomposition est identique à celle qui a été réalisée en analyse de la variance à un critère de classification. La seconde décomposition, qui constitue le ***modèle observé*** de l'analyse de la variance à deux critères de classification, fait apparaitre deux termes de **variation factorielle**, relatifs à l'un et l'autre des deux facteurs, un terme dit **d'interaction**, et un terme de **variation résiduelle.**

- Par **élévation au carré** et **sommation** pour les ***n p q*** **observations**, on obtient ensuite **l'*´equation d'analyse de la variance*** **:**

$$\boxed{\begin{aligned} \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n}(x_{ijk} - \bar{x}_{...})^2 &= q\,n\sum_{i=1}^{p}(\bar{x}_{i..} - \bar{x}_{...})^2 + p\,n\sum_{j=1}^{q}(\bar{x}_{.j.} - \bar{x}_{...})^2 \\ &+ n\sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x}_{...})^2 + \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n}(x_{ijk} - \bar{x}_{ij.})^2 . \end{aligned}}$$

Les deux premières composantes sont des **sommes de carrés d'écarts factorielles**, la troisième est une somme de carré d'écarts liée à **l'interaction**, et la quatrième est une **somme de carrés d'écarts résiduelle**.

$$\boxed{\mathrm{SCE}_t = \mathrm{SCE}_a + \mathrm{SCE}_b + \mathrm{SCE}_{ab} + \mathrm{SCE}_r}.$$

En affectant les lettres ***a* et *b*** , respectivement, à chacun des **deux critères de classification**, et en désignant les différents termes par **SCE*t* , SCE*a* , SCE*b* , SCE*ab*** et **SCE*r*** , on peut écrire aussi, de façon simplifiée :

- **Aux différentes sommes des carrés des écarts**, peuvent être associés des nombres de ***degrés de liberté***, qui sont liés par la relation :

$$\boxed{pqn-1=(p-1)+(q-1)+(p-1)(q-1)+pq(n-1)}.$$

p q n*−1 degrés de liberté** pour la **somme totale**, puisqu'elle fait intervenir globalement les ***p q n
observations individuelles,
p*−1 et *q*−1 degrés de liberté** pour les **deux sommes factorielles**, puisqu'elles sont calculées respectivement à partir de ***p et de ***q* moyennes**,
p q* (*n*−1) degrés de liberté** pour la **somme résiduelle**, puisqu'elle fait intervenir ***p q* échantillons** de ***n
observations, et
(*p* − 1) (*q* − 1) degrés de liberté pour la **somme des carrés des écarts de l'interaction**.

- Enfin, en **divisant** les différentes **sommes des carrés des écarts** par leurs **nombres de degrés de liberté**, on obtient les ***carrés moyens* CM*t* , CM*a* , CM*b* , CM*ab* et CM*r*** . L'ensemble des résultats peut alors être présent sous la forme d'un tableau **d'analyse de la variance ou ANOVA**

Sources de variation	Degrés de liberté	Sommes des carrés des écarts	Carrés moyens
Facteur a	$p-1$	SCE_a	CM_a
Facteur b	$q-1$	SCE_b	CM_b
Interaction	$(p-1)(q-1)$	SCE_{ab}	CM_{ab}
Variation résiduelle	$pq(n-1)$	SCE_r	CM_r
Totaux	$pqn-1$	SCE_t	

Tableau 3. Analyse de variance (ANOVA) à deux critères de classification : modèles croisés à Effectifs égaux

Application : Comparaison de **trois types de sondes** dans **deux types de sols** : **analyse de la variance**.

Au cours d'une étude relative aux problèmes d'échantillonnage du sol, on a comparé, dans **plusieurs types de sols**, différents **types de sondes** destinées à prélever des échantillons de terre, en effectuant chaque fois **diverses analyses chimiques**. On **s'intéresse** principalement aux **différences qui pourraient exister d'un type de sondes à l'autre** et aux **interférences éventuelles** des **types** de **sondes** avec les **types de sols**.

Le tableau suivant est relatif à **deux types de sols**, à **trois types de sondes**, et aux **teneurs en P2O5** , mg par 100 g de terre sèche, chacune des combinaisons **sol-sonde** ayant été l'objet de **quatre** prélèvements indépendants les uns des autres.

Ce tableau présente à la fois les données initiales ***xijk*** , et **les moyennes** par **type de sols** et **type de sondes** **¯*xij.*** , par **type de sols** **¯*xi..*** , par **type de sondes** **¯*x.j.*** , et générale **¯*x...*** , toutes les moyennes étant volontairement calculées avec une précision quelque peu abusive.

	Sonde 1 ($j = 1$)	Sonde 2 ($j = 2$)	Sonde 3 ($j = 3$)	
Sol 1 ($i = 1$)	43	41	42	
	45	42	44	
	46	43	46	
	53	44	48	
Sol 2 ($i = 2$)	40	35	37	
	40	37	39	
	40	40	40	
	43	40	40	
		$\bar{x}_{ij.}$		$\bar{x}_{i..}$
Sol 1	46,75	42,50	45,00	44,75
Sol 2	40,75	38,00	39,00	39,25
$\bar{x}_{.j.}$	43,75	40,25	42,00	42,00 ($\bar{x}_{...}$)

Tableau. Teneurs en P2O5 , en mg par 100 g de terre sèche, et moyennes observées, pour deux types de sols et trois types de sondes.

$\bar{x}_{ij.}$ = Moy. sonde1 pour le sol1=46,75= (43+45+46+53)/4
Moy. sonde1 pour le sol2=40,75=(40+40+40+43)/4

Moy. sonde2 pour le sol1=42,50= (41+42+43+44)/4 Moy. sonde2 pour le sol2=38,00= (35+37+40+40)/4 Moy. sonde3 pour le sol1=45,00= (42+44+46+48)/4 Moy. sonde3 pour le sol2=39,00= (37+39+40+40)/4

$\bar{x}_{.j.}$ =43,75= (46,75+40,75)/2 sonde1 pour les 2 sols
=40,25= (42,50+38,00)/2 sonde2 pour les 2 sols
=42,00= (45,00+39,00)/2 sonde3 pour les 2 sols

$\bar{x}_{i..}$ =44,75= (46,75+42,50+45,00)/3

=39,25= (40,75+38,00+39,00)/3

$$q\,n \sum_{i=1}^{p} (\bar{x}_{i..} - \bar{x}_{...})^2$$

$(\bar{x}_{...})$=42,00= (44,75+39,25)/2

Sources de variation	Degrés de liberté	Sommes des carrés des écarts	Carrés moyens
Types de sols	1	181,5	181,5
Types de sondes	2	49,0	24,5
Interaction	2	3,0	1,5
Variation résiduelle	18	112,5	6,25
Totaux	23	346,0	

Tableau. Comparaison de trois types de sondes dans deux types de sols : tableau partiel d'analyse de la variance (aspects descriptifs).

NB : Appliquer les formules et vérifier les valeurs des SCE et les CM

La réalisation de l'analyse de la variance

1◦ Nous désignerons par $X_{ij.}$ et SCE_{ij} , respectivement, les sommes et les sommes des carrés des écarts relatives aux différentes séries d'observations, et aussi par **X**$_{i..}$, ***X.j.*** et ***X...,*** les sommes relatives aux différentes modalités des deux critères de classification et la somme générale de toutes les observations.

2◦ Les différentes sommes des carrés des écarts peuvent alors être obtenues à l'aide des relations suivantes :

$$\text{SCE}_t = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n} x_{ijk}^2 - X_{...}^2/(p\,q\,n)\,, \quad \text{SCE}_a = \frac{1}{q\,n}\sum_{i=1}^{p} X_{i..}^2 - X_{...}^2/(p\,q\,n)\,,$$

$$\text{SCE}_b = \frac{1}{p\,n}\sum_{j=1}^{q} X_{.j.}^2 - X_{...}^2/(p\,q\,n) \quad \text{et} \quad \text{SCE}_r = \sum_{i=1}^{p}\sum_{j=1}^{q} \text{SCE}_{ij}\,,$$

Application. Comparaison de trois types de sondes dans deux types de sols : réalisation de l'analyse de la variance.

Les différentes sommes sont :

*X*11. =(43+45+46+53)= 187, *X*12. = 163 , *X*13. = 180 ,
*X*21. = 170 , *X*22. = 152 ,
*X*23. = 156 ,
*X*1.. = 537, *X*2.. = 471 , *X*.1. = 350 , *X*.2. = 322 , *X*.3. = 336 , *X*... = 1.008 .

Calculées selon les principes habituels, les sommes des carrés des écarts relatives aux six séries d'observations sont aussi :

$SCE11 = (43^2+45^2+46^2+53^2)-(43+45+46+53)^2/4=$ **56,8**

$SCE12 = (40^2+40^2+40^2+43^2)-(40+40+40+43)^2/4=$ **6,8**

$SCE13 = (42^2+44^2+46^2+48^2)-(42+44+46+48)^2/4=$ **20,0**

$SCE21 = (41^2+42^2+43^2+44^2)-(41+42+43+44)^2/4=$ **5**

$SCE22 = (35^2+37^2+40^2+40^2)-(35+37+40+40)^2/4=$ **18,0**

$SCE23 = (37^2+39^2+40^2+40^2)-(37+39+40+40)^2/4=$ **6,0**

SCEt $= (43-42)^2+(45-42)^2+(46-42)^2+(53-42^2+(40-42)^2+(40-42)^2+(40-42)^2+(43-42)^2+(42-42)^2+(44-42)^2+(46-42)^2+(48-42)^2+(41-42)^2+(42-42)^2+(43-42)^2+(44-42)^2+(35-42)^2+(37-42)^2+(40-42)^2+(40-42)^2+(37-42)^2+(39-42)^2+(40-42)^2+(40-42)^2=$ **346**

SCEr= SCE11+ SCE12+ SCE13+ SCE21 +SCE22 +SCE23= **56,8** +**6,8** + **20,0** +**5** +**18,0**+ **6,0= 112,6**

Comparaison de trois types de sondes dans deux types de sols : tableau complet d'analyse de la variance.

Sources de variation	Degrés de liberté	Sommes des carrés des écarts	Carrés moyens	F	P
Types de sols	1	181,5	181,5	29,0 ***	0,0000
Types de sondes	2	49,0	24,5	3,92 *	0,039
Interaction	2	3,0	1,5	0,24	0,79
Variation résiduelle	18	112,5	6,25		
Totaux	23	346,0			

On constate tout d'abord que l'interaction est non significative. Le test confirme donc la conclusion intuitive que les différences entre types de sondes ne dépendent pas des types de sols et vice versa Par contre, les différences observées sont très hautement significatives en ce qui concerne les types de sols, et juste significatives en ce qui concerne les types de sondes.

La comparaison, plus intéressante, des trois types de sondes est un problème qui peut être traité notamment par la méthode de **Newman et Keuls**. Cette méthode permet de montrer que les résultats obtenus à l'aide du premier type de sondes sont significativement supérieurs aux résultats fournis par le deuxième type de sondes, le troisième type conduisant à des résultats intermédiaires, qui ne sont pas significativement différents des deux autres.

les valeurs suivantes des variables *F* de Fisher-Snedecor : *Fa* = 1,69 ,
***Fb* = 0,66 et *Fab* = 2,96** les probabilités correspondantes sont
respectivement égales à **0,21 , 0,53 et 0,077**

-RESUME DE L'ANOVA A DEUX CRITERES DE CLASSIFICATION

• ANOVA à deux facteurs - Introduction

- **Définition**

  - Étude simultanée d'un facteur **A** à **p** modalités et d'un facteur **B** à **q** modalités.
  - Pour chaque couple de modalités (A, B) :
    - On a un échantillon
    - Tous les ***Eij*** sont de mêmes tailles ***n.***

- **Conditions d'applications de l'ANOVA**

- **Procédure de calcul d'une ANOVA**

  - Déterminer si les échantillons **varien**t 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 :

  - les variances sont homogènes
  - Au moins une des variances est différente des autres

→ Utilisation d'un **test de comparaison de plusieurs variances**

- **Conclusion**

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

- **Application :** Tests possibles

  - Influence du **facteur A seul**

- Influence du **facteur B seul**
- **Interaction** des **deux facteurs:**
    - ✓

Hypothèses

Si influence **d'un facteur sur la moyenne des populations est différente en l'absence ou en la présence de l'autre facteur**

$(H_0)_A$: Le facteur **A** n'a **pas d'influence sur la moyenne des populations**.

$(H_0)_B$: Le facteur **B** n'a **pas d'influence sur la moyenne des populations.**

H_1: Il n'y a **pas d'interaction entre les facteurs A et B.**

: Au moins **une des moyennes est différentes des autres.**

Variances totale, factorielle, résiduelle

Variance factorielle (variance intergroupe) :

- Dispersion des valeurs d'un échantillon à l'autre (influence du facteur) :

$$s_F^2 = \frac{n}{pq-1} \sum_{i=1}^{p} \sum_{j=1}^{q} \left(\bar{x}_{ij} - \bar{x}\right)^2$$

Variance résiduelle (variance intragroupe)

- Dispersion des valeurs à l'intérieur des échantillons (variabilité individuelle) :

$$s_R^2 = \frac{1}{(n-1)pq} \sum_{i=1}^{p} \sum_{j=1}^{q} (n-1) s_{ij}^2 = \frac{1}{pq} \sum_{i=1}^{p} \sum_{j=1}^{q} s_{ij}^2$$

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

Pour chaque échantillon E_{ij} de taille n , on calcule :

- moyenne

$$\bar{x}_{ij} = \frac{1}{n} \sum_{m=1}^{n} x_{ijm}$$

- variance expérimentale

$$s_{ij}^2 = \frac{1}{n-1} \sum_{m=1}^{n} \left(x_{ijm} - \bar{x}_{ij}\right)^2$$

Pour l'ensemble de l'expérience :

- Taille totale

$$N = npq$$

- Moyenne générale

$$\bar{x} = \frac{1}{npq} \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_{m=1}^{n} x_{ijm} = \frac{1}{npq} \sum_{i=1}^{p} \sum_{j=1}^{q} n \cdot \bar{x}_{ij} = \frac{1}{pq} \sum_{i=1}^{p} \sum_{j=1}^{q} \bar{x}_{ij}$$

- Variance totale

$$s^2 = \frac{1}{npq-1} \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_{m=1}^{n} \left(x_{ijm} - \bar{x}\right)^2$$

Décomposition de la variance factorielle

	A	B
Moyennes conditionnelles	$\bar{x}_{i.} = \frac{1}{q}\sum_{j=1}^{q} \bar{x}_{ij}$	$\bar{x}_{.j} = \frac{1}{p}\sum_{i=1}^{p} \bar{x}_{ij}$
Variance factorielle	$s^2_{FA} = \frac{qn}{p-1}\sum_{i=1}^{p}(\bar{x}_{i.} - \bar{\bar{x}})^2$	$s^2_{FB} = \frac{pn}{q-1}\sum_{j=1}^{q}(\bar{x}_{.j} - \bar{\bar{x}})^2$
	$s^2_{FAB} = \frac{n}{(p-1)(q-1)}\sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{x}_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}})^2$	

Tableau 33 : Décomposition de la variance factorielle

$SCE_F = SCE_{FA} + SCE_{FB} + SCE_{FAB}$

Conclusion

Tableau d'analyse de la variance :

Variation		SCE	ddl	CM	F
Factorielle		$SCE_F = (pq-1)s^2_F$	pq-1	$CM_F = s^2_F$	
	A	$SCE_{FA} = (p-1)s^2_{FA}$	p-1	$CM_{FA} = s^2_{FA}$	$F_A = \frac{s^2_{FA}}{s^2_R}$
	B	$SCE_{FB} = (q-1)s^2_{FB}$	q-1	$CM_{FB} = s^2_{FB}$	$F_B = \frac{s^2_{FB}}{s^2_R}$
	$\mathbf{A \times B}$	$SCE_{FAB} = (p-1)(q-1)s^2_{FAB}$	(p-1)(q-1)	$CM_{FAB} = s^2_{FAB}$	$F_{AB} = \frac{s^2_{FAB}}{s^2_R}$
Résiduelle		$SCE_R = (n-1)pq\, s^2_R$	(n-1)pq	$CM_R = s^2_R$	
Totale		$SCE_T = (npq-1)s^2$	nqp-1	$CM_T = s^2$	

$SCE_T = SCE_F + SCE_R$

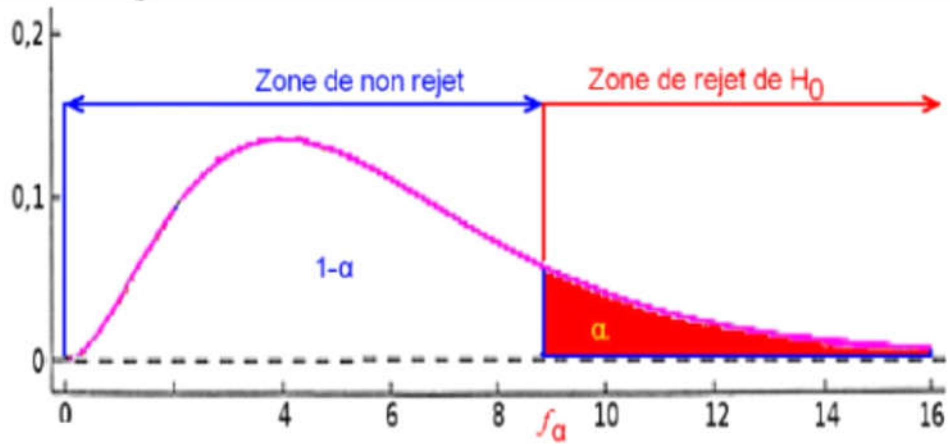
Sous l'hypothèse H0 :

- F_A suit une loi de Snédécor à $\nu_1 = p-1$ et $\nu_2 = npq-1$ ddl
- F_B suit une loi de Snédécor à $\nu_1 = q-1$ et $\nu_2 = npq-1$ ddl
- F_{AB} suit une loi de Snédécor à $\nu_1 = (p-1)(q-1)$ et $\nu_2 = npq-1$ ddl
- (tests unilatéraux : le rapport n'est pas obligatoirement supérieur à 1)

Choix du risque

- Risque de première espèce α (erreur commise lorsqu'on rejette H_0 à tort).

Décision



Zones de rejet de l'hypothèse nulle pour une distribution de Snédécor et un test unilatéral

- Si $F_A > f_{A\alpha}$ => **rejet de $(H_0)_A$ au risque** α :
  - La variance factorielle est significativement supérieure à la variance résiduelle : les moyennes diffèrent significativement entre-elles.

    → on attribue une influence significative au facteur A étudié.
- Si $F_B > f_{B\alpha}$ => **rejet de $(H_0)_B$ au risque** α :
  - La variance factorielle est significativement supérieure à la variance résiduelle : les moyennes diffèrent significativement entre-elles.

    → on attribue une influence significative au facteur B étudié.
- Si $F_{AB} > f_{AB\alpha}$ => **rejet de $(H_0)_{AB}$ au risque** α :
  - La variance factorielle est significativement supérieure à la variance résiduelle : les moyennes diffèrent significativement entre-elles.

    → il existe une interaction significative d'un facteur sur l'autre.
- Recherche du degré de signification p pour chaque test (recherche du risque α 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.**

Analyse De La Variance A Trois Critères De Classification

L'analyse de la variance à trois critères de classification :

Modèles croisés à effectifs égaux

Principes généraux

Nous présenterons successivement les *aspects* ***descriptifs*** et les *aspects inférentiels* de l'analyse à **trois critères**, en nous limitant, dans un premier temps, aux **échantillons de plusieurs observations**. Nous envisagerons ensuite le cas particulier des *échantillons d'une seule observation*.

- Les aspects descriptifs : échantillons de plusieurs observations

1◦ Pour trois critères de classification et dans le cas des effectifs égaux, on peut considérer qu'on a ***p q r*** échantillons ou séries d'observations d'effectif ***n***,et designer les observations individuelles par x_{ijkl} (*i* = 1 , . . . , *p* , *j* = 1 , . . . , *q* ,*k* = 1 , . . . , *r* , et *l* = 1 , . . . , *n*). A partir de ces observations, on peut calculer les différentes moyennes suivantes :

$$\bar{x}_{ijk.}\,,\quad \bar{x}_{ij..}\,,\quad \bar{x}_{i.k.}\,,\quad \bar{x}_{.jk.}\,,\quad \bar{x}_{i...}\,,\quad \bar{x}_{.j..}\,,\quad \bar{x}_{..k.}\quad \text{et}\quad \bar{x}_{....}\,.$$

Ces moyennes sont relatives, respectivement, aux différentes combinaisons des modalités des trois facteurs considérés simultanément **(*p q r* moyennes $\bar{x}_{ijk.}$),** aux différentes combinaisons des modalités des trois facteurs considérés deux à deux **(*p q* moyennes** $\bar{x}_{ij..}$, ***p r* moyennes** $\bar{x}_{i.k.}$, et

q r* moyennes),** aux différentes modalités des trois facteurs considérés individuellement **(*p* moyennes), *q* moyennes** $\bar{x}_{.j..}$ et ***r* moyennes** $\bar{x}_{..k.}$**)** et à l'ensemble des ***p q r n* observations (moyenne générale** $\bar{x}_{....}$ **).** Dans ces conditions, **le *modèle observé s'écrit :

$$
\begin{aligned}
x_{ijkl} - \bar{x}_{....} = \\
&(\bar{x}_{i...} - \bar{x}_{....}) + (\bar{x}_{.j..} - \bar{x}_{....}) + (\bar{x}_{..k.} - \bar{x}_{....}) + (\bar{x}_{ij..} - \bar{x}_{i...} - \bar{x}_{.j..} + \bar{x}_{....}) \\
&+ (\bar{x}_{i.k.} - \bar{x}_{i...} - \bar{x}_{..k.} + \bar{x}_{....}) + (\bar{x}_{.jk.} - \bar{x}_{.j..} - \bar{x}_{..k.} + \bar{x}_{....}) \\
&+ (\bar{x}_{ijk.} - \bar{x}_{ij..} - \bar{x}_{i.k.} - \bar{x}_{.jk.} + \bar{x}_{i...} + \bar{x}_{.j..} + \bar{x}_{..k.} - \bar{x}_{....}) + (x_{ijkl} - \bar{x}_{ijk.}).
\end{aligned}
$$

Le deuxième membre de ce modèle contient : **trois termes de variation factorielle** liés individuellement aux **trois critères de classification**, **trois termes d'interaction** des différents facteurs considérés deux à deux, **un terme d'interaction des trois facteurs** considérés simultanément, et **un terme de variation résiduelle**.

Les interactions des différents facteurs considérés deux à deux se présentent et doivent être interprétées comme en analyse de la variance à deux critères de classification. Ces interactions simples sont appelées ***interactions de deux facteurs*** ou ***interactions du premier ordre.***

L' ***´equation d'analyse de la variance*** relative à ce modèle observé est :

$$
\boxed{\mathrm{SCE}_t = \mathrm{SCE}_a + \mathrm{SCE}_b + \mathrm{SCE}_c + \mathrm{SCE}_{ab} + \mathrm{SCE}_{ac} + \mathrm{SCE}_{bc} + \mathrm{SCE}_{abc} + \mathrm{SCE}_r},
$$

Aux différentes sommes des carrés des écarts, correspondent des nombres de *degrés de liberté* liés par la relation :

$$
\begin{aligned}
p\,q\,r\,n - 1 = (p-1) + (q-1) + (r-1) + (p-1)(q-1) + (p-1)(r-1) \\
+ (q-1)(r-1) + (p-1)(q-1)(r-1) + p\,q\,r\,(n-1).
\end{aligned}
$$

La division des **sommes des carrés des écarts** par les **nombres de degrés de liberté** permet de définir les ***carrés moyens***, et l'ensemble des résultats peut être présente sous la forme d'un tableau d'analyse de la variance.

Application. ´Etude de la résistance de panneaux de particules à l'arrachage des clous : **réalisation de l'analyse de la variance**

Au cours d'un essai préliminaire, destin à préciser les conditions de mesure de cette propriété, on à étudié simultanément l'influence de trois facteurs : la grosseur des clous, le diamètre des anneaux sur lesquels sont déposées les éprouvettes soumises aux essais, et la vitesse d'arrachage. Les essais ont été effectues sur des éprouvettes carrées de 50 mm de coté, les modalités des trois facteurs étant : 6,5 et 8 mm de diamètre en ce qui concerne les tètes des clous (i = 1 et 2), 22 et 30 mm de diamètre en ce qui concerne les diamètres des anneaux servant de supports (j = 1 et 2), 22 , 45 et 90

mm par minute en ce qui concerne les vitesses d'arrachage (k = 1 , 2 et 3). En outre, cinq éprouvettes ont été utilisées pour chacune des 12 combinaisons des modalités des trois facteurs (l = 1 , . . . , 5).

i	j	k	l	x_{ijkl}	i	j	k	l	x_{ijkl}	i	j	k	l	x_{ijkl}	i	j	k	l	x_{ijkl}
1	1	1	1	54	1	2	1	1	54	2	1	1	1	67	2	2	1	1	67
1	1	1	2	56	1	2	1	2	51	2	1	1	2	71	2	2	1	2	66
1	1	1	3	58	1	2	1	3	47	2	1	1	3	72	2	2	1	3	62
1	1	1	4	51	1	2	1	4	51	2	1	1	4	70	2	2	1	4	71
1	1	1	5	57	1	2	1	5	59	2	1	1	5	81	2	2	1	5	69
1	1	2	1	57	1	2	2	1	52	2	1	2	1	79	2	2	2	1	72
1	1	2	2	58	1	2	2	2	56	2	1	2	2	80	2	2	2	2	67
1	1	2	3	61	1	2	2	3	52	2	1	2	3	81	2	2	2	3	75
1	1	2	4	59	1	2	2	4	52	2	1	2	4	80	2	2	2	4	70
1	1	2	5	55	1	2	2	5	53	2	1	2	5	85	2	2	2	5	71
1	1	3	1	60	1	2	3	1	63	2	1	3	1	78	2	2	3	1	78
1	1	3	2	68	1	2	3	2	54	2	1	3	2	78	2	2	3	2	81
1	1	3	3	61	1	2	3	3	65	2	1	3	3	77	2	2	3	3	67
1	1	3	4	61	1	2	3	4	62	2	1	3	4	83	2	2	3	4	76
1	1	3	5	67	1	2	3	5	60	2	1	3	5	79	2	2	3	5	75

Tableau **.Resistance de panneaux de particules à l'arrachage des clous (*xijkl*), en kg, pour deux grosseurs de clous (*i*), deux diamètres d'anneaux (*j*), trois vitesses d'arrachage (*k*), et dans chaque cas cinq éprouvettes (*l*).**

Resistance de panneaux de particules `a l'arrachage des clous (*xijkl*), en kg, pour deux grosseurs de clous (*i*), deux diamètres d'anneaux (*j*), deux vitesses d'arrachage (*k*), et dans chaque cas cinq éprouvettes.

Ce tableau permet de calculer facilement les sommes de produits suivantes :

$$\sum_{i,j,k,l} (i\, x_{ijkl}) = 363\,, \quad \sum_{i,j,k,l} (j\, x_{ijkl}) = -\,115\,, \quad \sum_{i,j,k,l} (k\, x_{ijkl}) = 81\,,$$

$$\sum_{i,j,k,l} (i\, j\, x_{ijkl}) = -\,37, \quad \sum_{i,j,k,l} (i\, k\, x_{ijkl}) = 47, \quad \sum_{i,j,k,l} (j\, k\, x_{ijkl}) = -\,35\,,$$

et

$$\sum_{i,j,k,l} (i\, j\, k\, x_{ijkl}) = -\,13\,,$$

$$\mathrm{SCE}_a = (363)^2/40 = 3.294{,}2\,, \quad \mathrm{SCE}_b = (-\,115)^2/40 = 330{,}6\,,$$

$$\mathrm{SCE}_c = (81)^2/40 = 164{,}0\,, \quad \mathrm{SCE}_{ab} = (-\,37)^2/40 = 34{,}2\,,$$

$$\mathrm{SCE}_{ac} = (47)^2/40 = 55{,}2\,, \quad \mathrm{SCE}_{bc} = (-\,35)^2/40 = 30{,}6\,,$$

et

$$\mathrm{SCE}_{abc} = (-\,13)^2/40 = 4{,}2\,.$$

Tableau . Etude de la résistance de panneaux de particules à l’arrachage des clous : **tableau d’analyse de la variance**

Sources de variation	Degrés de liberté	S. des carrés des écarts	Carrés moyens	F	P
Grosseurs des clous	1	3.294,2	3.294,2	297 ***	0,0000
Diamètres des anneaux	1	330,6	330,6	29,8 ***	0,0000
Vitesses d’arrachage	1	164,0	164,0	14,8 ***	0,0005
Clous-anneaux	1	34,2	34,2	3,09	0,088
Clous-vitesses	1	55,2	55,2	4,98 *	0,033
Anneaux-vitesses	1	30,6	30,6	2,76	0,11
Clous-anneaux-vitesses	1	4,2	4,2	0,38	0,54
Variation résiduelle	32	355,0	11,09		
Totaux	39	4.268,0			