Comparaison de Moyennes et ANOVA

Jean VAILLANT

Département de Mathématiques et Informatique, U.A.G.

27 Janvier 2012

Plan

Idée de base

Avoir une indication de l'ordre de grandeur d'une série de valeurs mesurées (ou observées)

- $$\frac{\text{Total des valeurs observées}}{\text{Nombre de valeurs observées}}$$

notée traditionnellement $\bar{x}$.

Idée de base

Avoir une indication de l'ordre de grandeur d'une série de valeurs mesurées (ou observées)

- $$\frac{\text{Total des valeurs observées}}{\text{Nombre de valeurs observées}}$$

  notée traditionnellement $\bar{x}$.
- Question sous jacente : par quelle valeur unique pourrait-on remplacer toutes les valeurs pour avoir la même somme?

Idée de base

- Exemple : moyenne des évaluations dans 2 collèges d'effectifs 400 et 600 respectivement.

  Collège 1 : $x_{1;1}, \cdots, x_{1;400}$

  Collège 2 : $x_{2;1}, \cdots, x_{2;600}$

$$\bar{x} = \frac{(x_{1;1} + \cdots + x_{1;400}) + (x_{2;1} + \cdots + x_{2;600})}{400 + 600}.$$

La moyenne des évaluations pour chaque collège :
$\bar{x}_1 = 12,4$ et $\bar{x}_2 = 13,8$.

Idée de base

- Exemple : moyenne des évaluations dans 2 collèges d'effectifs 400 et 600 respectivement.

  Collège 1 : $x_{1;1}, \cdots, x_{1;400}$

  Collège 2 : $x_{2;1}, \cdots, x_{2;600}$

$$\bar{x} = \frac{(x_{1;1} + \cdots + x_{1;400}) + (x_{2;1} + \cdots + x_{2;600})}{400 + 600}.$$

  La moyenne des évaluations pour chaque collège :
  $\bar{x}_1 = 12,4$ et $\bar{x}_2 = 13,8$.

- Que faire si données perdues sauf moyennes et effectifs?
  $\bar{x} = \frac{1}{2}(\bar{x}_1 + \bar{x}_2) = 13,1$ (INCORRECT)

$$\bar{x} = \frac{400\bar{x}_1 + 600\bar{x}_2}{400 + 600} = 0,4\bar{x}_1 + 0,6\bar{x}_2 = 13,24 \quad \text{(CORRECT)}.$$

Moyennes arithmétiques

Formellement : n nombres $x_1, x_2, \cdots, x_n$. On distingue les moyennes suivantes :

- arithmétique simple $\frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$

Moyennes arithmétiques

Formellement : n nombres $x_1, x_2, \cdots, x_n$. On distingue les moyennes suivantes :

- arithmétique simple $\frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$
- arithmétique pondérée $p_1 x_1 + p_2 x_2 + \cdots + p_n x_n = \sum_{i=1}^{n} p_i x_i$,

  les poids $p_1, p_2, \cdots, p_n$ étant positifs et tels que $\sum_{i=1}^{n} p_i = 1$.

Moyenne géométrique

- géométrique $(x_1 x_2 \cdots x_n)^{\frac{1}{n}} = (\prod_{i=1}^{n} x_i)^{\frac{1}{n}}$, les $x_i > 0$

Exemple : Quotient des effectifs d'un collège entre 2009 et 2011

Années	2008	2009	2010	2011
Effectifs	420	448	430	479
Quotient	-	1,07	0,96	1,12

Moyenne géométrique :
$(1,07 \times 0,96 \times 1,12)^{\frac{1}{3}} = 1,15^{\frac{1}{3}} = 1,045$.

Moyenne géométrique

- géométrique $(x_1 x_2 \cdots x_n)^{\frac{1}{n}} = (\prod_{i=1}^{n} x_i)^{\frac{1}{n}}$, les $x_i > 0$

  Exemple : Quotient des effectifs d'un collège entre 2009 et 2011

Années	2008	2009	2010	2011
Effectifs	420	448	430	479
Quotient	-	1,07	0,96	1,12

  Moyenne géométrique :
  $(1,07 \times 0,96 \times 1,12)^{\frac{1}{3}} = 1,15^{\frac{1}{3}} = 1,045$.

- Question sous jacente : par quelle valeur unique pourrait-on remplacer toutes les valeurs pour avoir le même produit?
  $\bar{x} = \frac{1}{3}(1,07 + 0,96 + 1,12) = 1,050$ (INCORRECT)

  $(1,07 \times 0,96 \times 1,12)^{\frac{1}{3}} = 1,15^{\frac{1}{3}} = 1,045$ (CORRECT).

Moyenne harmonique

- harmonique $\dfrac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}} = \dfrac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$, les $x_i > 0$.

Moyenne harmonique

- harmonique $\dfrac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}} = \dfrac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$, les $x_i > 0$.
- Exemple : élève faisant le trajet domicile-lycée à la vitesse constante de $2km/h$ à l'aller et $4km/h$ au retour.

  Vitesse moyenne du trajet aller-retour?
  $\bar{x} = \frac{1}{2}(2+4) = 3,000$ (INCORRECT)

  $\dfrac{2}{\frac{1}{2}+\frac{1}{4}} = 2,667$ (CORRECT).

Moyenne harmonique

- harmonique $\frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$, les $x_i > 0$.
- Exemple : élève faisant le trajet domicile-lycée à la vitesse constante de $2km/h$ à l'aller et $4km/h$ au retour.

  Vitesse moyenne du trajet aller-retour?
  $\bar{x} = \frac{1}{2}(2+4) = 3,000$ (INCORRECT)

  $\frac{2}{\frac{1}{2}+\frac{1}{4}} = 2,667$ (CORRECT).
- Question sous jacente : par quelle valeur unique pourrait-on remplacer toutes les valeurs pour avoir la même somme des inverses?

Liens entre les 3 types de moyenne, f-moyenne

Soit f fonction continue strictement monotone sur l'intervalle $I \subset \mathbb{R}$.
Soit $x_1, \cdots, x_n$ une série de valeurs de I.

Il existe un unique élément m de I tel que $f(m) = \frac{1}{n} \sum_{i=1}^{n} f(x_i)$.

Liens entre les 3 types de moyenne, f-moyenne

Soit f fonction continue strictement monotone sur l'intervalle $I \subset \mathbb{R}$.
Soit $x_1, \cdots, x_n$ une série de valeurs de I.

Il existe un unique élément m de I tel que $f(m) = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$.

Preuve existence:

$\min_{i=1,\cdots,n} f(x_i) \leq \frac{1}{n}\sum_{i=1}^{n} f(x_i) \leq \max_{i=1,\cdots,n} f(x_i)$ donc $\frac{1}{n}\sum_{i=1}^{n} f(x_i) \in f(I)$.

Preuve unicité: f est injective.

Liens entre les 3 types de moyenne, f-moyenne

Soit f fonction continue strictement monotone sur l'intervalle $I \subset \mathbb{R}$.
Soit $x_1, \cdots, x_n$ une série de valeurs de I.

Il existe un unique élément m de I tel que $f(m) = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$.

Preuve existence:

$\min_{i=1,\cdots,n} f(x_i) \leq \frac{1}{n}\sum_{i=1}^{n} f(x_i) \leq \max_{i=1,\cdots,n} f(x_i)$ donc $\frac{1}{n}\sum_{i=1}^{n} f(x_i) \in f(I)$.

Preuve unicité: f est injective.

m est appelé f-moyenne des $x_1, \cdots, x_n$ et l'on a $m = f^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} f(x_i)\right)$.

Liens entre les 3 types de moyenne, f-moyenne

Soit f fonction continue strictement monotone sur l'intervalle $I \subset \mathbb{R}$.
Soit $x_1, \cdots, x_n$ une série de valeurs de I.

Il existe un unique élément m de I tel que $f(m) = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$.

Preuve existence:

$\min_{i=1,\cdots,n} f(x_i) \leq \frac{1}{n}\sum_{i=1}^{n} f(x_i) \leq \max_{i=1,\cdots,n} f(x_i)$ donc $\frac{1}{n}\sum_{i=1}^{n} f(x_i) \in f(I)$.

Preuve unicité: f est injective.

m est appelé f-moyenne des $x_1, \cdots, x_n$ et l'on a $m = f^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} f(x_i)\right)$.

Question sous jacente : par quelle valeur unique pourrait-on remplacer tous les x_i pour avoir la même somme des $f(x_i)$?

Si $f(x) = x$, la moyenne est arithmétique
Si $f(x) = Ln(x)$, la moyenne est géométrique
Si $f(x) = 1/x$, la moyenne est harmonique.

Positions respectives des f-moyennes

Si f strictement convexe et décroissante, alors $m < \bar{x}$.
Ainsi, pour $I = \mathbb{R}_+^*$, moyenne harmonique $<$ moyenne arithmétique.

Si f strictement concave et croissante, alors $m < \bar{x}$.
Ainsi, pour $I = \mathbb{R}_+^*$, moyenne géométrique $<$ moyenne arithmétique.

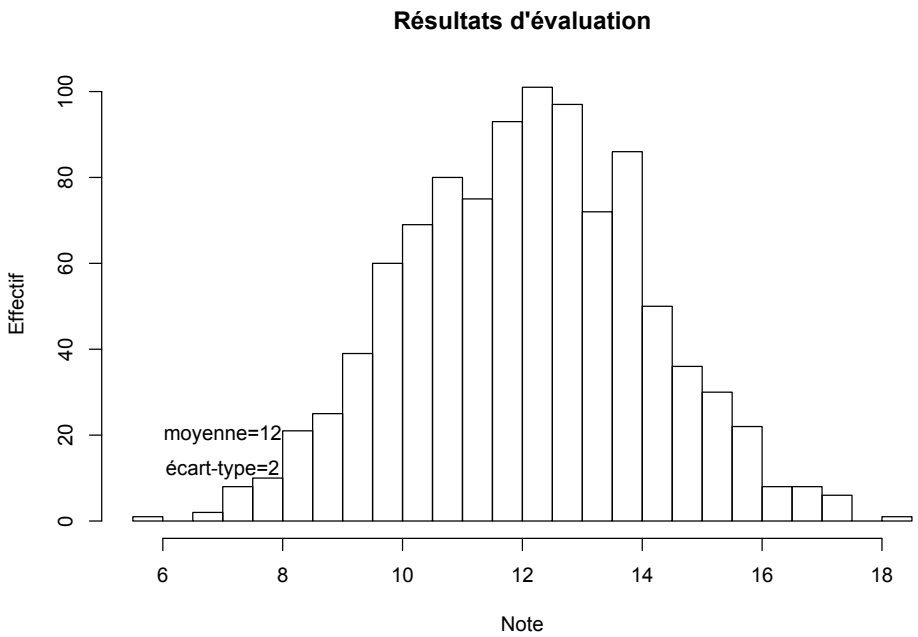
En utilisant la convexité de $x \mapsto Ln(1/x)$, on démontre que
moyenne harmonique $<$ géométrique.

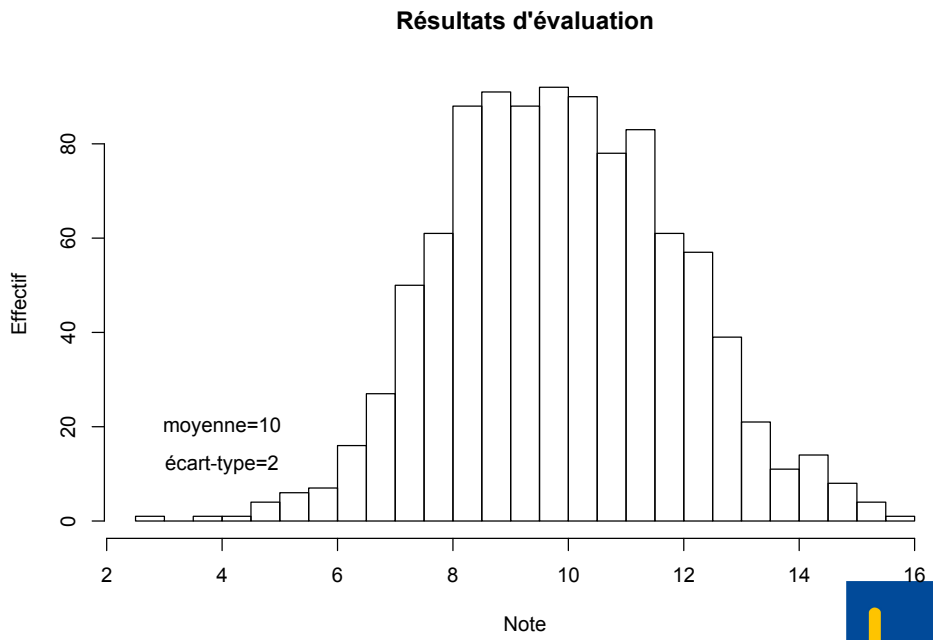
CONCLUSIONS :
Soit $x_1, \cdots, x_n$ des valeurs non toutes identiques dans $\mathbb{R}_+^*$, alors

$$\min_{i=1,\cdots,n} x_i \quad \leq \quad \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \quad \leq \quad \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} \quad \leq \quad \bar{x} \quad \leq \max_{i=1,\cdots,n} x_i \quad .$$

Investigation par histogrammes

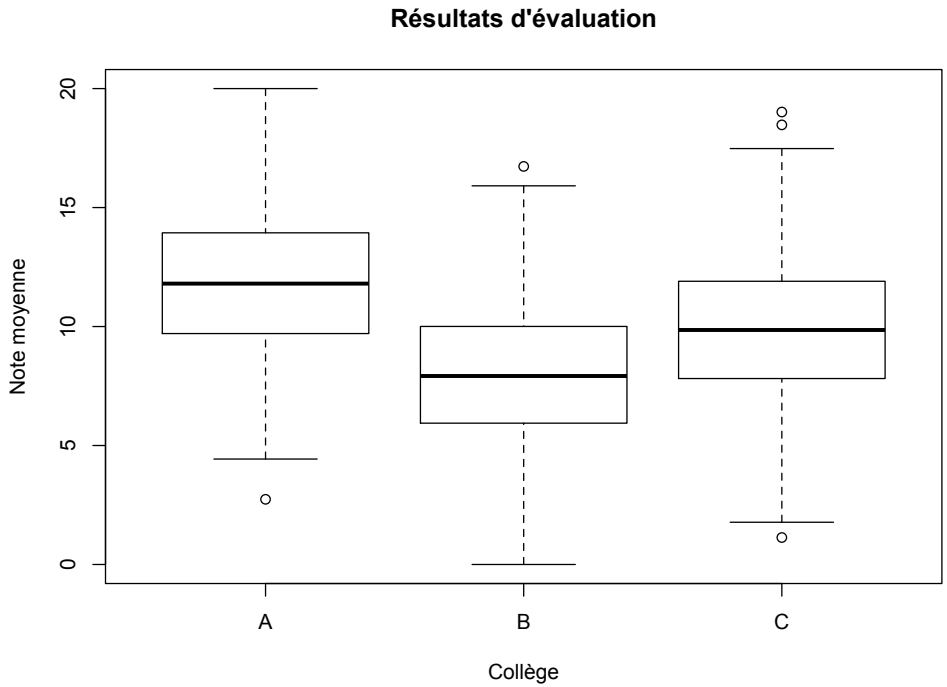
Contexte : Une variable quantitative Y à expliquer par une ou plusieurs variables qualitative(s)
Visualiser : la distribution de Y pour chaque modalité des variables qualitatives.





Investigation par Box-plot

Visualiser : la distribution de Y pour chaque modalité des variables qualitatives.



Contexte de l'ANOVA

ANOVA (ANalysis Of VAriance) = méthode de comparaison de groupes.

Contexte : Une variable quantitative à expliquer par une ou plusieurs variables qualitative(s) appelée(s) facteur(s).

- Question: $\exists$? influence des variables qualitatives sur la variable quantitative ?

  Exemple : résultats des évaluations d'un collège à l'autre?

Contexte de l'ANOVA

ANOVA (ANalysis Of VAriance) = méthode de comparaison de groupes.

Contexte : Une variable quantitative à expliquer par une ou plusieurs variables qualitative(s) appelée(s) facteur(s).

- Question: $\exists$? influence des variables qualitatives sur la variable quantitative ?

  Exemple : résultats des évaluations d'un collège à l'autre?

- Remarque : si la variable qualitative n'a que deux modalités, test de Student

  Exemple : Résultats chez les garçons et chez les filles

Contexte de l'ANOVA

ANOVA (ANalysis Of VAriance) = méthode de comparaison de groupes.

Contexte : Une variable quantitative à expliquer par une ou plusieurs variables qualitative(s) appelée(s) facteur(s).

- Question: $\exists$? influence des variables qualitatives sur la variable quantitative ?

  Exemple : résultats des évaluations d'un collège à l'autre?

- Remarque : si la variable qualitative n'a que deux modalités, test de Student

  Exemple : Résultats chez les garçons et chez les filles

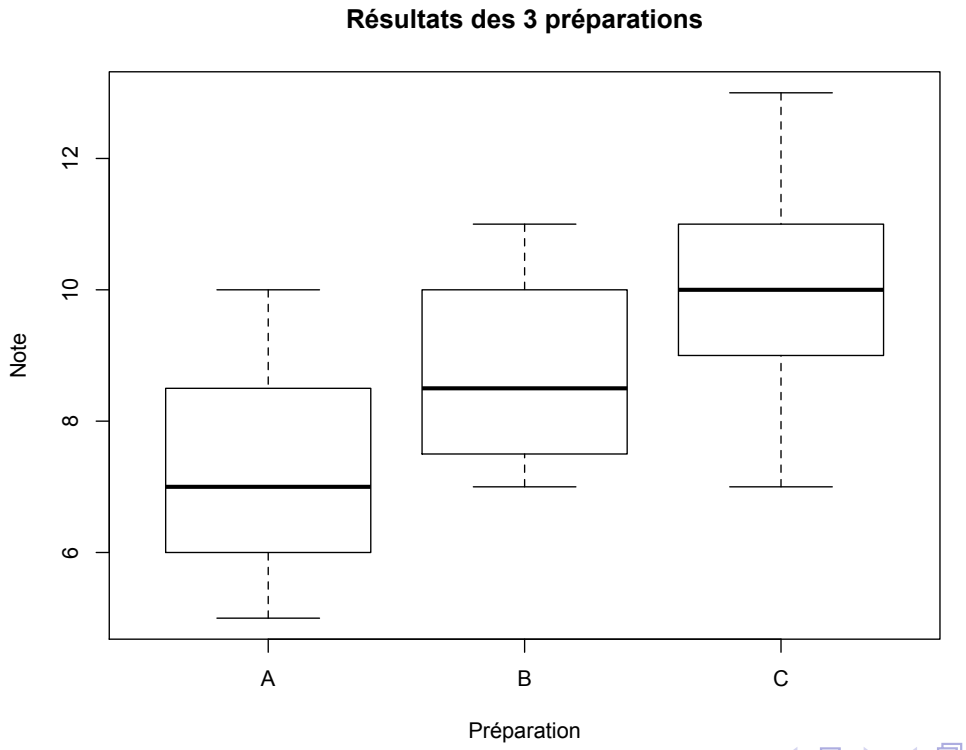
- L'ANOVA permet d'étudier plus de 2 modalités

Exemple

Notes de 25 élèves subissant trois préparations distinctes
Préparation A: 5,6,6,7,7,8,9,10 –> moyenne 7.25
Préparation B: 7,7,8,8,9,10,10,11 –> moyenne 8.75
Préparation C: 7,9,9,10,10,10,11,12,13 –> moyenne 10.11
Ces différences sont-elles significatives?

Exemple

Notes de 25 élèves subissant trois préparations distinctes
Préparation A: 5,6,6,7,7,8,9,10 –> moyenne 7.25
Préparation B: 7,7,8,8,9,10,10,11 –> moyenne 8.75
Préparation C: 7,9,9,10,10,10,11,12,13 –> moyenne 10.11
Ces différences sont-elles significatives?



Ecarts significatifs entre groupes?

Cela va dépendre des :
- Moyennes de groupe
- Ecart-types de groupe
- Tailles de groupe

Outils : tests statistiques d'hypothèses

Le Problème de test est :

H_0 : "les moyennes attendues sont égales pour tous les groupes"

contre

H_1 : "les moyennes attendues ne sont pas égales pour tous les groupes".

Si H_1 retenue, pousser l'investigation pour savoir quel(s) groupe(s) diffère(nt) des autres

Suppositions de l'ANOVA

- Les observations suivent approximativement la loi Normale

  - vérification par visualisation d'histogrammes et test de normalité affaiblie à l'unimodalité

  si existence de valeurs extrêmes, préférer le test de Kruskall-Wallis (basé sur les médianes)

Suppositions de l'ANOVA

- Les observations suivent approximativement la loi Normale

  - vérification par visualisation d'histogrammes et test de normalité affaiblie à l'unimodalité

  si existence de valeurs extrêmes, préférer le test de Kruskall-Wallis (basé sur les médianes)

- Ecart-types approximativement égaux d'un groupe à l'autre

  Règle empirique : rapport entre plus grand ecart-type et plus petit ≤ 2.

Vérification des suppositions

Préparation	Taille	Moyenne	Médiane	Ecart-type
A	8	7,250	7,000	1,669
B	8	8,75	8,500	1,488
C	9	10,111	10,000	1,764

Plus grand écart-type : 1,764
Plus petit écart-type : 1,488
Règle empirique : $1{,}488 \times 2 = 2{,}976 > 1{,}764$

Notations de l'ANOVA à un facteur

n = nombre d'unités statistiques (ou individus statistiques)
I = nombre de groupes = nombre de modalités du facteur
n_i = effectif du groupe i
x_{ij} = valeur observées sur l'unité j du groupe i
$\bar{x}_i$= moyenne du groupe i
$\bar{x}$= moyenne globale (c'est-à-dire tous groupes confondus)
s_i= écart-type du groupe i
avec

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

Principe de l'ANOVA à un facteur

L'ANOVA à un facteur, mesure deux sources de variation :

- la variation intra-groupe

  en considérant l'écart entre chaque valeur et la moyenne de son groupe

$$SCR = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 = \sum_{i=1}^{I} n_i s_i^2$$

  SCR= Somme des Carrés Résiduelle (ou intra-groupe).

Principe de l'ANOVA à un facteur

L'ANOVA à un facteur, mesure deux sources de variation :

- la variation intra-groupe

  en considérant l'écart entre chaque valeur et la moyenne de son groupe

$$SCR = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 = \sum_{i=1}^{I} n_i s_i^2$$

  SCR= Somme des Carrés Résiduelle (ou intra-groupe).

- la variation inter-groupe

  en considérant l'écart entre moyennes de groupe et moyenne globale

$$\sum_{i=1}^{I} \sum_{j=1}^{n_i} (x_i - \bar{x})^2 = \sum_{i=1}^{I} n_i (x_i - \bar{x})^2$$

  SCF= Somme des Carrés Factorielle (ou inter-groupe).

Théorème d'ANOVA à un facteur

$SCT =$ Somme des Carrés due à la variation Totale.

$$SCT = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$$

prend en compte l'écart entre chaque valeur et la moyenne globale.

Théorème d'ANOVA : $SCT = SCF + SCR$.
basé sur la décomposition $x_{ij} - \bar{x} = \bar{x}_i - \bar{x} + x_{ij} - \bar{x}_i$

Modèle sous-jacent :

$x_{ij} = \mu + \alpha_i + e_{ij}$ avec les e_{ij} indépendants de loi $\mathcal{N}(0, \sigma^2)$.

Principe de compensation des effets : $\sum_{i=1}^{I} \alpha_i = 0$.

Autre énoncé du modèle :

Valeur observée= valeur attendue tous groupes confondus + effet du groupe d'appartenance + erreur résiduelle.

Tableau d'ANOVA à un facteur

Source de variation	Somme des carrés	Degrés de liberté	Carré moyen	F de Fisher
Totale	SCT	$n-1$	$\text{CMT}=\frac{SCT}{n-1}$	
Factorielle	SCF	$I-1$	$CMF=\frac{SCF}{I-1}$	$F=\frac{CMF}{CMR}$
Résiduelle	SCR	$n-I$	$CMR=\frac{SCR}{n-I}$	

Formalisation du problème de test; Règle décisionnelle

x_{ij} réalisation d'une variable aléatoire X_{ij} avec $E(X_{ij}) = \mu_i = \mu + \alpha_i$.
Problème de test :
$H_0 : \forall i \in [\![1, I]\!], \mu_i = \mu$ contre $H_1 : \exists i \in [\![1, I]\!], \mu_i \neq \mu$.
ou encore
$H_0 : \forall i \in [\![1, I]\!], \alpha_i = 0$ contre $H_1 : \exists i \in [\![1, I]\!], \alpha_i \neq 0$.
Sous H_0, F suit la loi de Fisher-Snedecor à $I-1$ et $n-I$ degrés de liberté.
Règle de décision au niveau de signification α :

- Rejet de H_0 si $F > f_{I-1,n-I,1-\alpha}$
- Non rejet de H_0 si $F \leq f_{I-1,n-I,1-\alpha}$

où $f_{I-1,n-I,1-\alpha}$ fractile d'ordre $1-\alpha$ de la loi de Fisher-Snedecor à $I-1$ et $n-I$ degrés de liberté.
Règle de décision basée sur la p**-value** p :

- Rejet de H_0 si $p < \alpha$
- Non rejet de H_0 si $p \geq \alpha$.

Objets et commandes sous l'environnement R

Appel à la fonction "lm" (linear model) qui effectue le traitement statistique du modèle linéaire sous-jacent à l'ANOVA.

Appel à la fonction "aov" (analysis of variance) qui calcule le tableau d'ANOVA et effectue le test de Fisher.

Appel à la fonction "summary" qui effectue un résumé des calculs effectués par "lm" et "aov" .

L'objet Evaluation est le vecteur correspondant à la série statistique des évaluations

L'objet Préparation est le vecteur correspondant à la série statistique des préparations subies.

Lignes de commande :
> summary(lm(Evaluation~Preparation))
> summary(aov(Evaluation~Preparation))

Résultats sous l'environnement R

Paramètre	Estimation	Ecart-type	t de Student	p-value
α_2	1,5000	0,8250	1,818	0,08266
α_3	2,8611	0,8017	3,569	0,00172 **

Signification et p-value : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Source de Variation	Somme des carrés	Degrés de liberté	Carré moyen	Statistique de Fisher	p-value
Factorielle	34,671	2	17,3356	6,3682	0,00658 **
Résiduelle	59,889	22	2,7222		

Signification et p-value : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

L'écart intuitif entre préparations est confirmé par les tests statistiques.
Ce n'est pas toujours le cas!!!

MERCI DE VOTRE ATTENTION !