Université Lumière Lyon 2
UFR de Sciences Economiques et de Gestion
M2 - Chargé d'Etudes Economiques

ANOVA - Support de cours (3)

Rafik Abdesselam
rafik.abdesselam@univ-lyon2.fr
http://perso.univ-lyon2.fr/~rabdesse/Documents/

Année Universitaire 2024 - 2025

ANOVA

M2 APE : Economic and Statistical Studies
Data Science

Science des Données - Analyse des données

Plan du cours

Objectif du cours

UFR de Sciences économiques et de Gestion

UNIVERSITÉ DE LYON

- Présenter les concepts et les conditions d'application de l'analyse de la variance.
- L'analyse de variance abrégée sous le terme anglais ANOVA (ANalysis Of VAriance) est une technique statistique permettant de comparer les moyennes d'un nombre quelconque de populations, contrairement à ce que pourrait laisser penser son nom.
- Son lien avec la régression est présenté, mais d'une façon générale, elle consiste à comparer plusieurs moyennes d'échantillons provenant de populations normales.
- Une large place est accordée dans ce cours aux exemples et exercices sur données réelles traités avec les logiciels SAS et SPAD.

Objectif du cours

- ANOVA
  - Pré-requis : Quelques notions de Statistique & Probabilités.
  - Approche pédagogique : une séance de cours magistral et (1) séance de travaux dirigés. Supports informatique : Logiciels SPAD - SAS.
  - Matériel pédagogique : (1) Polycopié de support de cours, (1) polycopié de travaux dirigés ainsi que de nombreux fichiers de données réelles (Tables SAS - version 9.2. - Bases de données SPAD).
  - Quelques références bibliographiques :
    [1] Dagnelie, P. Analyse statistique à plusieurs variables. Gembloux, Presses agronomiques, 1986, 362p.
    [2] Mervyn, G.Marasinghe, William J. Kennedy, SAS for Data Analysis, Intermediate Statistical Methods. Statistics and Computing, Springer, 2008.
    [3] Saporta, G. Probabilités Analyse des données et Statistique. Editions technip, 1990.

Introduction

- Les techniques d'analyse de variance ANOVA sont des outils entrant dans le cadre du Modèle Linéaire Général et où une variable quantitative est expliquée par une ou plusieurs variables qualitatives.
- L'objectif est alors de comparer les moyennes empiriques de la variable quantitative observée pour les différentes catégories d'unités statistiques.
- Ces catégories sont définies par l'observation des variables qualitatives ou facteurs prenant différentes modalités ou encore de variables quantitatives découpées en classes ou niveaux.
- Une combinaison de niveaux définit une cellule, groupe ou traitement.

Introduction

- L'analyse de variance ANOVA est une technique statistique de tests et d'estimation qui permet d'analyser l'effet d'une voire plusieurs variables qualitatives sur une variable continue.
- L'ANOVA est très utilisée dans le contexte des plans d'expériences et des traitements des données expérimentales.
- D'un point de vue modélisation, l'ANOVA n'est autre qu'une régression multiple sur variables explicatives qualitatives (nominales).
- Les principes essentiels de l'analyse de la variance à un et à deux facteurs de classification avec/sans interaction seront exposés.

Terminologie

- Les variables qualitatives sont appelées "facteurs ou critères" et leurs modalités "niveaux" du facteur. En présence de plusieurs facteurs, une combinaison de niveaux est un "traitement".
- Statistiquement, l'ANOVA est une généralisation du test de Student pour comparer plus de 2 moyennes.
- Il arrive fréquemment que les données soient groupées en classes selon certains critères ou facteurs tels que, par exemple, l'âge, l'appartenance sociale, la région géographique, etc.

Exemple introductif

- Prenons comme exemple, le cas d'une étude sur la fréquence d'utilisation des moyens de transports en commun.
- On peut supposer que celle-ci sera différente en fonction de l'âge des personnes interrogées.
- Il est donc naturel de diviser la population en classes d'âges (par exemple : adolescents, adultes, personnes âgées) avant d'effectuer l'échantillonnage.
- Sur la base des observations des différents échantillons constitués, la question sera de savoir s'il existe effectivement une différence significative d'utilisation des transports en commun entre les classes d'usagers considérées.
- Ceci revient à effectuer un test de comparaison multiple de moyennes.

Exemple illustratif

- Les données représentent la fréquence journalière d'utilisation des moyens de transports en commun de trois groupes d'usagers.

Table: Fréquence d'utilisation des moyens de transports.

	Adolescents	Adultes	Personnes âgées
	3	5	3
	6	7	3
	5	6	2
	6	7	2
	5	5	5
Moyenne	5	6	3

- Le problème consiste à détecter les différences, si elles existent, entre les moyennes des populations à partir desquelles ces observations ont été obtenues.
- Comparer la différence entre les moyennes des groupes d'usagers, mesurée en terme de variabilité, tout en tenant compte de la variabilité existant entre les usagers à l'intérieur de chaque groupe.

Exemple : Cas particulier 1

- Pour bien distinguer entre ces deux types de variabilité, considérons les données des deux tableaux fictifs suivants.

Table: Exemple 1 de variation nulle à l'intérieur.

	Adolescents	Adultes	Personnes âgées
	5	6	3
	5	6	3
	5	6	3
	5	6	3
	5	6	3
Moyenne	5	6	3

- Toutes les observations dans chaque échantillon ont ici la même valeur. Il n'y a donc aucune variation à l'intérieur des groupes (ou échantillons d'usagers), mais il y a une variation entre les groupes d'usagers, puisque les moyennes d'échantillonnage sont différentes.

Exemple : Cas particulier 2

Table: Exemple 2 de variation nulle entre les groupes.

	Adolescents	Adultes	Personnes âgées
	8	4	3
	6	5	7
	5	9	7
	4	7	8
	7	5	5
Moyenne	6	6	6

- Par contre, dans ce cas, la moyenne de chaque groupe d'usagers est identique. Il n'y a donc pas de variation entre les groupes, mais il y a une variation à l'intérieur des groupes puisque toutes les observations dans chaque groupe n'ont pas la même valeur.
- En pratique, les observations obtenues ne sont ni exactement identiques, ni de moyennes égales ; elles sont hétérogènes comme les données de l'exemple introductif.

ANOVA à un facteur contrôlé

- **Les données** :

• n observations réparties dans p groupes ou échantillons.

• Chaque échantillon j (j = 1,p) contient n_j observations, correspondant à un niveau différent d'un facteur. Les tailles n_j des échantillons pouvant être égales ou différentes, $n = \sum_j^p n_j$.

- Le tableau suivant illustre un exemple de données.

Ech.1	Ech.2	...	Ech.j	...	Ech.p
x_{11}	x_{12}	...	x_{1j}	...	x_{1p}
x_{21}	⋮	⋮	⋮	⋮	⋮
x_{i1}	x_{i2}	...	⋮	⋮	x_{ip}
⋮	$x_{n_{22}2}$	⋮	⋮	⋮	⋮
⋮		⋮	$x_{n_{jj}j}$	⋮	⋮
$x_{n_1 1}$		...			⋮
		...			$x_{n_{pp}p}$
$\overline{x}_1$	$\overline{x}_2$	...	$\overline{x}_j$	...	$\overline{x}_p$

• $\overline{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j}$ la moyenne des n_j observations du groupe j.

• $\overline{x} = \frac{\sum_{j=1}^{p}\sum_{i=1}^{n_j} x_{ij}}{n} = \frac{\sum_{j=1}^{p} n_j \overline{x}_j}{n}$ la moyenne globale.

ANOVA à un facteur

- Le facteur à p niveaux est supposé avoir un effet uniquement sur les moyennes des distributions et non sur leur variance. Il s'agit d'un test de comparaison de p moyennes.
- D'une façon générale, il s'agit de tester s'il existe une différence "significative" entre les moyennes m_j (j = 1,p) des p populations dans lesquelles ont été prélevés les p échantillons indépendants de taille n_j (j = 1,p) de l'étude.
- En d'autres termes, effectuer le test statistique suivant :

$$\begin{cases} H_0 : m_1 = m_2 = ... = m_p & \text{pas de différence significative.} \\ H_1 : m_j \neq m_k \quad j \neq k & \text{différence significative.} \end{cases}$$

- Il suffit donc qu'une moyenne soit différente de toutes les autres pour que l'hypothèse nulle H_0 soit rejetée.
- Il s'agit là d'une généralisation à p populations du test classique (t de Student) de comparaison de moyennes de 2 échantillons.

Condition d'application - Ecarts

- Conditions d'application à vérifier avant toute utilisation d'une analyse de la variance :
  1. Les échantillons doivent être indépendants.
  2. Les distributions des populations considérées doivent être normales (hypothèse de normalité - test paramétrique).
  3. Les populations d'où sont prélevés les échantillons doivent posséder la même variance : $\sigma^2 = \sigma_1^2 = \sigma_2^2 = ... = \sigma_p^2$ (hypothèse d'homocédasticité).
- Pour procéder à une analyse de la variance, on s'intéresse à trois types d'écart :
  1. Chaque observation par rapport à sa moyenne respective : $(x_{ij} - \overline{x}_j)$.
  2. Chaque moyenne d'échantillonnage par rapport à la moyenne globale : $(\overline{x}_j - \overline{x})$.
  3. Chaque observation par rapport à la moyenne globale : $(x_{ij} - \overline{x})$.

Tableau d'analyse de la variance à 1 facteur

- Les résultats obtenus sont résumés dans un tableau d'analyse de la variance (ou tableau ANOVA).

Sources de variation	***Somme des carrés***	***Degrés de liberté***	***Carrés moyens***	**F**
Entre Inter / Between	SC_{ent} $\sum_{j=1}^{p} n_j(\overline{x}_j - \overline{x})^2$	p - 1	$CM_{ent} = s^2_{ent}$ $\frac{SC_{ent}}{p-1}$	$\frac{s^2_{ent}}{s^2_{int}}$
Intérieur Intra / Within	SC_{int} $\sum_{j=1}^{p}\sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2$	n - p	$CM_{int} = s^2_{int}$ $\frac{SC_{int}}{n-p}$	
Totale Total	SC_{tot} $\sum_{j=1}^{p}\sum_{i=1}^{n} (x_{ij} - \overline{x})^2$	n - 1		

- La variation totale est la somme de 2 variations : $SC_{tot} = SC_{ent} + SC_{int}$.
- Cette propriété montre pourquoi la technique de comparaison de moyennes est appelée analyse de la variance, car ces sommes de carrés sont utilisées pour estimer des variances.
- En effet, le test réalisé consiste à décomposer la variance (constante) de x en deux parties : une variance interclasse ($CM_{ent} = s^2_{ent}$) et une variance intraclasse ou erreur ($CM_{int} = s^2_{int}$) puis à établir le test de Fisher (rapport de 2 variances $F = \frac{s^2_{ent}}{s^2_{int}}$).

ANOVA à 1 facteur contrôlé

Approche régression :

- On peut associé à l'analyse de la variance à 1 facteur le modèle de régression linéaire multiple suivant :

$$y_{ij} = \mu + \alpha_j + \varepsilon_{ij}, \qquad i = 1, ..., n_j \text{ et } j = 1, ..., p.$$

où, y_{ij} désigne la ième observation du jème échantillon,
μ est la moyenne générale commune,
α_j est l'effet du niveau jème du facteur,
ε_{ij} est l'erreur relative à l'observation y_{ij}.

- L'objectif est de tester certains paramètres du modèle notamment l'hypothèse nulle suivante :

$$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_p$$

ce qui signifie que les p facteurs ont un effet identique.
L'hypothèse alternative étant formulée comme suit :

$$H_1 : \text{ au moins une différence } \alpha_j \neq \alpha_{j'} \quad j \neq j'$$

c'est-à-dire que les α_j ne sont pas tous identiques.

Exemple d'application

- Les données représentent le niveau de production de 27 cadres employés, affectés à une tâche d'assemblage et de vérification, selon leur statut dans l'entreprise.

Table: Niveau de Production selon le Statut - Niveau de responsabilité de l'employé.

	Junior	Intermédiaire	Supérieur
	45 49 45 45 48 45 47 47 46	51 51 49 50 46 49 50 46 51 48 49	49 50 52 48 48 51 51
Effectif	9	11	7
Moyenne	46.333	49.091	49.857

Conditions d'application

- Effet du Statut de l'employé sur le niveau de Production.

1. Hypothèse de normalité :

Table: The Univariate Procedure: Tests for Normality.

	Test		Statistic		p-Value
Junior	Shapiro-Wilk	W	0.851103	$Pr < W$	0.0766 ✓
Intermédiaire	Shapiro-Wilk	W	0.869325	$Pr < W$	0.0760 ✓
Supérieur	Shapiro-Wilk	W	0.913363	$Pr < W$	0.4197 ✓

2. Hypothèse d'homoscédasticité - Egalité des variances :

Table: The GLM Procedure: Bartlett's Test for Homogeneity of Production Variance.

Source	DF	Chi-Square	$Pr > ChiSq$
Statut	2	0.3344	0.8461 ✓

Principaux résultats - SAS

Table: The MEANS Procedure: Analysis Variable : Production.

Statut	N	Mean	Std Dev	Minimum	Maximum
Junior	9	46.3333333	1.5000000	45.0000	49.0000
Intermédiaire	11	49.0909091	1.8140863	46.0000	51.0000
Supérieur	7	49.8571429	1.5735916	48.0000	52.0000

Table: The ANOVA Procedure: Effet du Statut sur le niveau de production.

Source	***Sum of squares***	***DF***	***Mean square***	***F-value***	**Pr > F**
Model	58.5300625	2	29.2650313	10.68	0.0005 ✓
Error	65.7662338	24	2.7402597		
Total	124.2962963	26			

Principaux résultats - SAS

- Localiser les disparités

Table: Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	Statut
A	49.8571	7	Supérieur
A	49.0909	11	Intermédiaire
B	46.3333	9	Junior

Table: Comparisons significant at the 0.05 level are indicated by ***.

Statut Comparison	Difference Between Means	95% Confidence Limits	
S - I	0.7662	-0.8856	2.4181	
S - J	3.5238	1.8020	5.2456	***
I - S	-0.7662	-2.4181	0.8856	
I - J	2.7576	1.2220	4.2932	***
J - S	-3.5238	-5.2456	-1.8020	***
J - I	-2.7576	-4.2932	-1.2220	***

ANOVA à deux facteurs

- Supposons maintenant le cas où les n observations de l'échantillon sont classées selon 2 facteurs A et B respectivement à p et q modalités-niveaux.
- Les n observations peuvent être réparties dans un tableau à p lignes (facteur A) et q colonnes (facteur B).
- Les trois questions principales que l'on se pose lors d'une analyse de variance à 2 facteurs :
  1. Y a-t-il un effet du facteur A : les moyennes mesurées sur les p populations définies par le facteur A sont-elles différentes ?
  2. Y a-t-il un effet du facteur B : les moyennes mesurées sur les q populations définies par le facteur B sont-elles différentes ?
  3. Y a-t-il un effet conjugué des deux facteurs A et B : une interaction entre les moyennes du facteur A et celles du facteur B ?

- Les résultats obtenus sont résumés dans un tableau d'analyse de la variance.

Sources de variation	***Somme carrés***	***Degrés de liberté***	***Carrés moyens***	**F**
Facteur A	SC_{entA}	p - 1	$CM_{entA} = \frac{SC_{entA}}{p-1}$	$\frac{CM_{entA}}{CM_{int}}$
Facteur B	SC_{entB}	q - 1	$CM_{entB} = \frac{SC_{entB}}{q-1}$	$\frac{CM_{entB}}{CM_{int}}$
Interaction AB	SC_{entAB}	(p - 1)(q - 1)	$CM_{entAB} = \frac{SC_{entAB}}{(p-1)(q-1)}$	$\frac{CM_{entAB}}{CM_{int}}$
Intérieur	SC_{int}	n - pq	$CM_{int} = \frac{SC_{int}}{n-pq}$	
Total	SC_{tot}	n - 1		

Exemple d'application

- Les données représentent le niveau de production de 27 cadres employés, affectés à une tâche d'assemblage et de vérification, selon leur statut - niveau de responsabilité et leur Sexe - genre dans l'entreprise.
- SPAD : Répartition de la production selon le statut et le Sexe de l'employé.

Effectif Moyenne Ecart-type	Féminin	Masculin	Ensemble
Intermédiaire	9 49.778 1.030	2 46.000 0.000	11 49.091 1.730
Junior	2 48.500 0.500	7 45.714 0.881	9 46.333 1.414
Supérieur	5 50.400 1.356	2 48.500 0.500	7 49.857 1.457
Ensemble	16 49.813 1.236	11 46.273 1.286	27 48.370 2.146

Conditions d'application

- Effet du Sexe de l'employé sur le niveau de Production.

1. Hypothèse de normalité :

Table: The Univariate Procedure: Tests for Normality.

	Test		Statistic		p-Value
Féminin	Shapiro-Wilk	W	0.869325	$Pr < W$	0.0974 ✓
Masculin	Shapiro-Wilk	W	0.851103	$Pr < W$	0.0736 ✓

2. Hypothèse d'homoscédasticité - Egalité des variances :

Table: The ANOVA Procedure: Levene's Test for Homogeneity of Production Variance.

Source	DF	Sum of Squares	Mean Square	F-Value	$Pr > F$
Sexe	1	0.1027	0.1027	0.03	0.8546 ✓
Error	25	74.8838	2.9954		

Principaux résultats - SAS

- Effet du Statut et du Sexe de l'employé sur la production.

1. Significativité du modèle dans son ensemble.

Source	*Sum of squares*	*DF*	*Mean square*	*F-value*	Pr > F
Model	99.1121693	5	19.8224339	16.53	<.0001 ✓
Error	25.1841270	21	1.1992441		
Total	124.2962963	26			

2. Significativité des facteurs et de leur interaction.

Source	*Type III SS*	*DF*	*Mean square*	*F-value*	Pr > F
SEXE	36.65908991	1	36.65908991	30.57	<.0001 ✓
STATUT	16.83835896	2	8.41917948	7.02	0.0046 ✓
SEXE*STATUT	2.70344328	2	1.35172164	1.13	0.3428

Principaux résultats - SPAD

- Effet du Statut et du Sexe de l'employé sur la production.

- Significativité des niveaux des facteurs.

Iden Libellé CRITERE(S)	Coeff.	E.type	T	Proba.	V.test
FEMI - Féminin	1.4106	0.255	5.529	0.000 ✓	4.30
MASC - Masculin	-1.4106	0.255	5.529	0.000 ✓	-4.30
INTE - Intermédiaire	-0.2598	0.355	0.731	0.473	-0.72
JUNI - Junior	-1.0415	0.360	2.896	0.009 ✓	-2.63
SUPE - Supérieur	1.3013	0.367	3.541	0.002 ✓	3.10

Principaux résultats - SPAD

- Effet du Statut et du Sexe de l'employé sur la production.

- Significativité de l'interaction des niveaux des facteurs.

Iden Libellé Interaction(s)	Coeff.	E.Type	T	Proba.	V.test
FEMI - Féminin INTE - Intermédiaire	0.4783	0.355	1.347	0.192	1.30
FEMI - Féminin JUNI - Junior	-0.0177	0.360	0.049	0.961	-0.05
FEMI - Féminin SUPE - Supérieur	-0.4606	0.363	1.268	0.219	-1.23
MASC - Masculin INTE - Intermédiaire	-0.4783	0.355	1.347	0.192	-1.30
MASC - Masculin JUNI - Junior	0.0177	0.360	0.049	0.961	0.05
MASC - Masculin SUPE - Supérieur	0.4606	0.363	1.268	0.219	1.23
Constante	48.1487	0.255	188.722	0.000	12.42

ANOVA à 2 facteurs contrôlés

Approche régression :

- On peut associé à l'analyse de la variance à 2 facteurs le modèle de régression suivant :

$$y_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon_{ijk},$$
$$i = 1, ..., m \ ; \ j = 1, ..., p \text{ et } k = 1, ..., q.$$

- Les trois questions principales évoquées précédemment se traduisent de la façon suivante :

1. $H_0 : \alpha_1 = \alpha_2 = ... = \alpha_p$
2. $H_0 : \beta_1 = \beta_2 = ... = \beta_q$
3. $H_0 : \gamma_{11} = \gamma_{12} = ... = \gamma_{jk} = ... = \gamma_{pq}$

- Modèle de régression à $(1 + p + q + pq)$ paramètres à estimer.

Instructions sous SAS & SPAD

- Le tableau suivant résume la syntaxe des procédures et instructions SAS nécessaires pour obtenir les principaux résultats de l'ANOVA - ANCOVA.

Hypothèse	Test	Instruction SAS	Procédure
Normalité	histogramme	nom-var / normal	univariate
Normalité	Shapiro-Wilk	normal	univariate
Normalité	qq-plot	nom-var / normal	univariate
Normalité	pp-plot	nom-var / normal	capability
Homoscédasticité	Bartlett ou Levene	means / hovtest =	anova ou glm
ANOVA 1 facteur			anova ou glm
ANOVA 2 facteurs			lmg
Comparaisons		means / t cldiff	anova ou glm
Disparités		means / tukey lines	anova ou glm

- Instructions d'un projet SPAD pour l'ANOVA - ANCOVA.

Groupe de méthodes	Méthode
Statistiques descriptives	Tests statistiques - Test de normalité (Shapiro-Wilk)
Scoring et modélisation	Vareg (Régression, analyse de variance-covariance)