

MANOVA

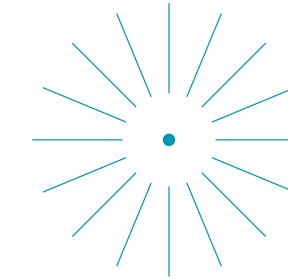
MULTIVARIATE ANALYSIS OF VARIANCE

BAUDAT Timéo, BOUCHIBTI Yasmine, GRIMAJ Meryem

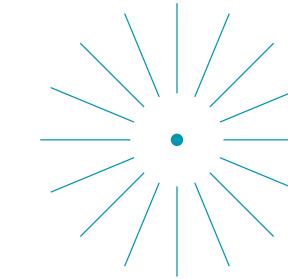
07 novembre 2024

Problématique et Objectifs

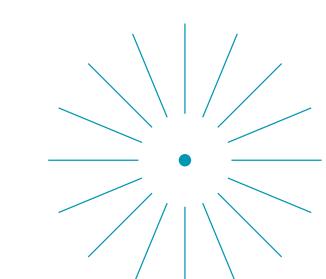
de la MANOVA



Comparer les moyennes de plusieurs groupes pour **plusieurs variables répondantes dépendantes**

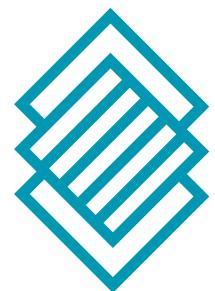


Tester si les facteurs (ou variables indépendantes) ont un **effet significatif** sur un ensemble de variables réponses dépendantes combinées.



Tenir compte des **corrélations** entre les variables réponses dépendantes.

Données pour la MANOVA



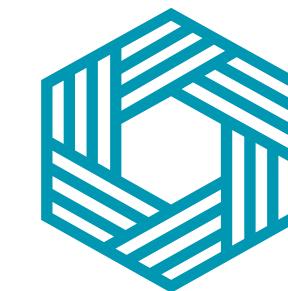
Les variables
réponses
dépendantes sont
quantitatives



Les variables
explicatives
indépendantes sont
qualitatives



Les variables
réponses sont ni trop
faiblement ni trop
fortement corrélées
($0,3 < r < 0,9$)



Les variables
réponses sont des
caractéristiques d'un
même thème

Exemple de données

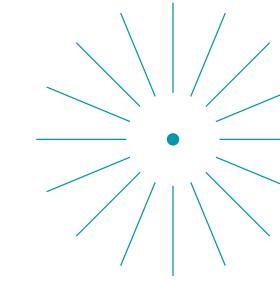
*Le jeu de données **swiss** est inclus dans R
Contient des **informations socio-économiques**
sur 47 districts de la Suisse en 1888.*



Variables réponses

Fertility

Taux de fertilité (nombre moyen d'enfants par femme).



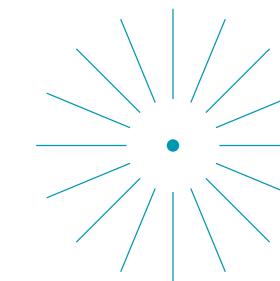
Infant.Mortality

Taux de mortalité infantile (nombre de décès d'enfants de moins d'un an par 1 000 naissances vivantes)

Variables explicatives

Examination

Le pourcentage de jeunes hommes ayant réussi le test d'aptitude militaire.
Réparties en 2 classes



Catholic

Pourcentage de la population catholique.
Réparties en 2 classes

Education

Taux d'éducation, mesurant l'accès à l'éducation.
Réparties en 3 classes



Agriculture

Pourcentage de la population active travaillant dans l'agriculture.
Réparties en 2 classes

Conditions d'applications de la MANOVA

- 1) Variables réponses **quantitatives continues.** (Y_i)
- 2) Variables explicatives **qualitatives.** (X)
- 3) Observations **indépendantes** dans le même groupes et d'un groupe à l'autre.
- 4) **Grande taille de l'échantillon.**
- 5) Absence de **valeurs aberrantes** univariées ou multivariées -> affaiblir l'hypothèse de normalité multivariée.



Conditions d'applications de la MANOVA

- 6) **Absence d'une trop forte multicolinéarité** entre les variables cibles (des ANOVA prises séparément si les corrélations trop élevées supérieures à 0.9).
- 7) **Normalité multivariée** (“mvnormtest” ou “ mshapiro.test()” sur R).
- 8) **Homoscédasticité**



Ecriture du modèle

Comme l'ANOVA, La MANOVA est basés sur le modèle linéaire général :

$$Y = X\beta + \epsilon$$

- Y est la matrice des variables dépendantes de dimension **n×p**
- X est la matrice des variables indépendantes de dimension **n×m**
- β est la matrice des coefficients de régression de dimension **m×p**
- ϵ est la matrice des résidus de dimension **n×p**

- Les tests sont décrits en termes de matrices de somme des carrés et des produits des écarts B et W.
- La matrice B (Between) est appelée la matrice d'hypothèses
- La matrice W (Within) est appelée la matrice d'erreur.
- On obtient ainsi par sommation la matrice totale T

$$T = W + B$$

$$\text{SPEtotal} = \text{SPEmodèle} + \text{SPErésidu}$$

Hypothèse nulle

Dans le cas simple d'une MANOVA à un seul facteur fixe à g modalités (une seule variable indépendante) et p variables dépendantes, l'hypothèse nulle H0 d'absence de différences significatives entre les g moyennes des populations s'écrit :

$$\mu_{11} = \mu_{21} = \dots = \mu_{g1}$$

$$\mu_{12} = \mu_{22} = \dots = \mu_{g2}$$

.....

$$\mu_{1p} = \mu_{2p} = \dots = \mu_{gp}$$

μ_{ij} : Moyenne théorique de la population i relative à la variable j

H0 peut s'écrire encore sous la forme :

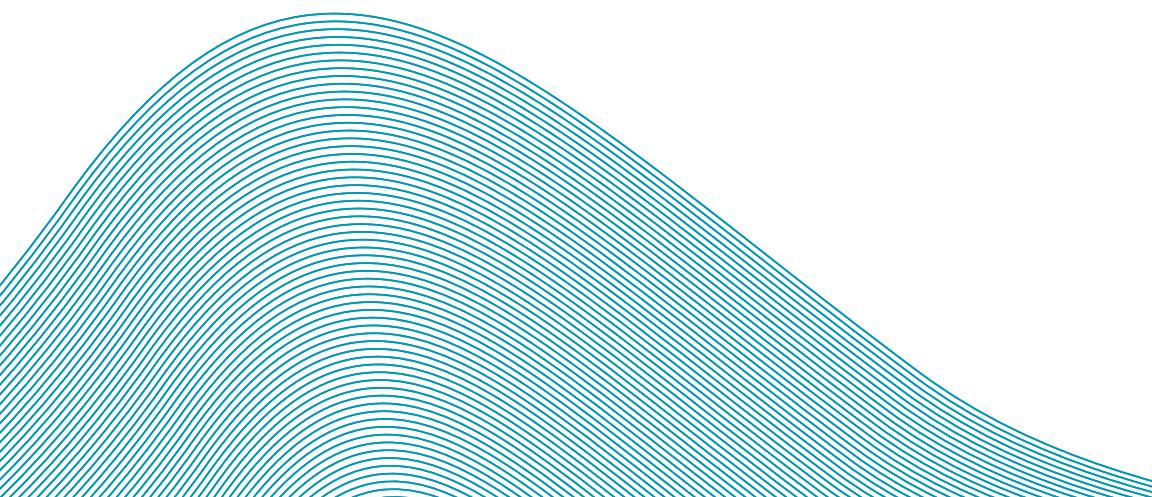
$$H_0: \mu_1 = \mu_2 = \dots = \mu_g$$

Ecriture du modèle

$$\begin{pmatrix} Fertility \\ Infant. Mortality \end{pmatrix} = \beta_0 + \beta_1 \times \text{Agriculture} + \beta_2 \times \text{Education} + \beta_3 \times \text{Catholic} + \beta_4 \times \text{Examination} + \epsilon$$

Test statistiques

- On se souvient que dans le cas de l'ANOVA on utilise la statistique F de Fisher pour tester statistiquement les effets principaux et les interactions.
- Pour la MANOVA, on dispose de plusieurs tests statistiques :
 - Le lambda de Wilks (Λ)
 - La trace de Lawley-Hotelling
 - La trace de Pillai**
 - Le test de la plus grande racine de Roy



La trace de Pillai

$$V = \text{trace}(B(W + B)^{-1})$$

$$F_{obs} = \frac{(2k_2 + s + 1)}{(2k_1 + s + 1)} \frac{V}{s - V}$$

- $s = \min(p, Vh)$, qui suit approximativement une distribution F de Fisher à $s(2k_1 + s + 1)$ et $s(2k_2 + s + 1)$ degrés de liberté.
- k_1 et k_2 sont des coefficients ajustés pour que F_{obs} suive une loi de Fisher pour tester l'hypothèse statistique.
- C'est un test en général réalisé au seuil de signification alpha=5%

Interprétation des données

```
> summary(result)
```

	Df	Pillai	approx F	num Df	den Df	Pr (>F)	
Agriculture	1	0.11451	2.5864	2	40	0.087832	.
Education	2	0.56165	8.0050	4	82	1.689e-05	***
Catholic	1	0.25983	7.0209	2	40	0.002436	**
Examination	1	0.12191	2.7767	2	40	0.074267	.
Residuals	41						

Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05 '.' 0.1 ' '
							1

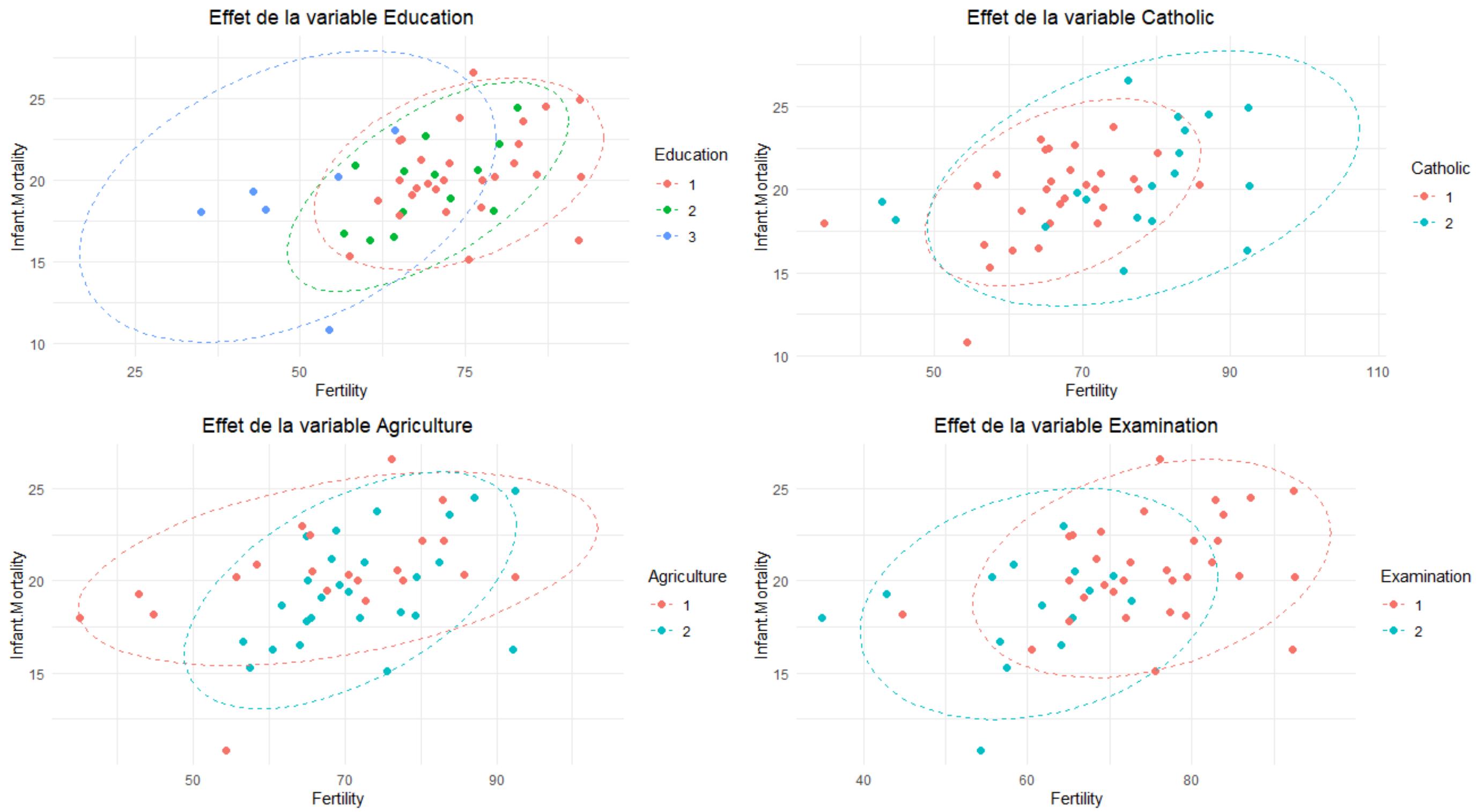
Package MANOVA.RM

On utilise la trace de Pillai.

Ici on regarde les effets **globaux**.

Ici 2 effets sont significatifs (p-value > 0.05). L'effet d'Examination et d'Agriculture ne sont pas significatifs.

Représentation graphique



Test post-hoc

2 manières les plus courantes :

- ANOVA univariée avec correction de Bonferroni
- **Analyse discriminante**

```
> lda_fit <- lda(Education ~ Fertility + Infant.Mortality, data = swiss)
> print(lda_fit)
Call:
lda(Education ~ Fertility + Infant.Mortality, data = swiss)
```

Prior probabilities of groups:

	1	2	3
0.5957447	0.2765957	0.1276596	

Group means:

	Fertility	Infant.Mortality
1	74.92143	20.41786
2	69.38462	19.70000
3	49.48333	18.25000

Coefficients of linear discriminants:

	LD1	LD2
Fertility	-0.10779230	-0.03172269
Infant.Mortality	0.02622322	0.36915775

Proportion of trace:

LD1	LD2
0.9978	0.0022

Comparaison avec ANOVA

ANOVA

MANOVA

Objectif principal

Tester des différences de moyennes entre plusieurs groupes **pour une seule variable dépendante.**

Nombre de variables réponses

Une seule variable.

Hypothèses

Les résidus sont normalement distribués et l'homogénéité des variances est respectée.

Type de résultats

Valeur p, **F-statistique** pour évaluer l'effet du facteur.

Tests post-hoc

Nécessaires si le test ANOVA est significatif pour déterminer quelles moyennes sont différentes.

MANOVA

Tester des différences de moyennes entre plusieurs groupes **pour plusieurs variables dépendantes simultanément.**

Plusieurs variables dépendantes.

Les résidus **pour chaque variable dépendante** sont normalement distribués, et l'homogénéité des covariances doit être respectée.

Valeur p, **statistiques multivariées** (comme le lambda de Wilks ou la trace de Pillai) utilisées pour évaluer l'effet des facteurs sur les variables dépendantes.

Peut nécessiter des analyses post-hoc spécifiques pour chaque variable dépendante si la MANOVA est significative.
(ANOVA univariée, Analyse discriminante).

Comparaison avec ANOVA

Conclusion