

Démarche statistique

Premiers pas avec

David Causeur

L'Institut Agro

IRMAR CNRS UMR 6625

<https://dcauseur.netlify.app>



Plan

- 1 Effet à l'échelle d'une population
- 2 Décider à partir de données
- 3 Effet 'groupe'
 - Comparaison de groupes
 - Analyse de variance à un facteur
 - Estimation des paramètres d'effet
 - Test de Fisher
 - Le cas particulier de la comparaison de 2 groupes
 - Décrire un effet groupe
 - Test avec des données appariées
- 4 Effet linéaire

Plan

- 1 Effet à l'échelle d'une population
- 2 Décider à partir de données
- 3 Effet 'groupe'
 - Comparaison de groupes
 - Analyse de variance à un facteur
 - Estimation des paramètres d'effet
 - Test de Fisher
 - Le cas particulier de la comparaison de 2 groupes
 - Décrire un effet groupe
 - Test avec des données appariées
- 4 Effet linéaire

Plan

- 1 Effet à l'échelle d'une population
- 2 Décider à partir de données
- 3 Effet 'groupe'
 - Comparaison de groupes
 - Analyse de variance à un facteur
 - Estimation des paramètres d'effet
 - Test de Fisher
 - Le cas particulier de la comparaison de 2 groupes
 - Décrire un effet groupe
 - Test avec des données appariées
- 4 Effet linéaire

Distribution d'une variable quantitative

'**X a un effet sur Y**'

peut être résumé à

'**la valeur moyenne de Y parmi les individus d'un même groupe diffère selon le groupe**'.

Hypothèse de normalité par groupe

Si Y_{ij} est la valeur de la variable réponse pour le j ème individu ($j = 1, \dots, n_i$) du i ème groupe ($i = 1, \dots, I$) :

$$Y_{ij} = \mu_i + \varepsilon_{ij},$$

où $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma)$ est l'erreur résiduelle.

Il y a un effet de X sur Y si, pour au moins deux groupes $i \neq i'$, $\mu_i \neq \mu_{i'}$.

Modèle statistique pour un effet 'groupe'

Deux composantes dans la décomposition $Y_{ij} = \mu_i + \varepsilon_{ij}$:

- μ_i pour les variations dues au facteur ;
⇒ les μ_i sont / **paramètres inconnus**.
- ε_{ij} pour les variations aléatoires intra-groupes.
⇒ σ , l'**écart-type résiduel**, est un autre paramètre inconnu.

Test d'un effet

Test de l'effet de X sur Y :

$$\begin{cases} H_0 : \mu_1 = \dots = \mu_I = \mu \text{ (pas d'effet de } X) \\ H_1 : \text{Pour au moins un couple } (i, i'), \text{ avec } i \neq i', \mu_i \neq \mu_{i'}. \end{cases}$$

Choix entre deux modèles :

- le **modèle nul** (X n'a pas d'effet sur Y) [sous-modèle]

$$Y_{ij} = \mu + \varepsilon_{ij}.$$

- et le **modèle non-nul** :

$$Y_{ij} = \mu_i + \varepsilon_{ij}.$$

Ajustement du modèle

Ajuster un modèle revient à **estimer** ses paramètres, c-à-d leur donner une valeur de telle manière que le modèle soit aussi *proche* que possible des données.

Proche ? ... selon le **critère des moindres carrés** :

$$SS(\mu_1, \dots, \mu_I) = \sum_{j=1}^{n_1} (Y_{1j} - \mu_1)^2 + \dots + \sum_{j=1}^{n_I} (Y_{Ij} - \mu_I)^2.$$

En minimisant séparément les termes $\sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2$:

$$\hat{\mu}_i = \frac{Y_{i1} + \dots + Y_{in_i}}{n_i} = \bar{Y}_{i\bullet}$$

Ajustement du modèle

Un **estimateur** $\hat{\theta}$ de θ est une fonction des données, garantissant que $\hat{\theta}$ est *proche* de θ .

... On appelle $\hat{\mu}_i$ l'**estimateur des moindres carrés** de μ_i .

Ajustement du modèle

$\hat{\mu}_i$ est-il proche de μ_i ?

$$\hat{\mu}_i - \mu_i = \frac{(Y_{i1} - \mu_i) + \dots + (Y_{in_i} - \mu_i)}{n_i} = \frac{\varepsilon_{i1} + \dots + \varepsilon_{in_i}}{n_i}.$$

- $\mathbb{E}(\hat{\mu}_i - \mu_i) = 0$: on dit que $\hat{\mu}_i$ est **non-biaisé** ;
- $\text{Var}(\hat{\mu}_i - \mu_i)$ a pour expression :

$$\text{Var}(\hat{\mu}_i - \mu_i) = \frac{\text{Var}(\varepsilon_{i1}) + \dots + \text{Var}(\varepsilon_{in_i})}{n_i^2} = \frac{\sigma^2 + \dots + \sigma^2}{n_i^2} = \frac{\sigma^2}{n_i}.$$

- $\hat{\mu}_i - \mu_i$ est distribué selon une loi normale.

En résumé, $\hat{\mu}_i - \mu_i \sim \mathcal{N}(0; \frac{\sigma}{\sqrt{n_i}})$

Ajustement du modèle

► Ajustement du modèle dans \mathbb{R}

Le **modèle d'analyse de la variance à un facteur** pour un effet groupe sur Y est le suivant :

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0; \sigma),$$

où $\alpha_2, \dots, \alpha_I$ sont les paramètres d'effet.

Écart-type résiduel

Erreurs d'ajustement ou résidus : $\hat{\varepsilon}_{ij} = Y_{ij} - \hat{\mu}_i = Y_{ij} - \bar{Y}_{i\bullet}$.

Comme σ^2 est la variance de l'erreur résiduelle :

$$\begin{aligned}\hat{\sigma}^2 &= \frac{\sum_{j=1}^{n_1} (Y_{1j} - \bar{Y}_{1\bullet})^2 + \dots + \sum_{j=1}^{n_l} (Y_{lj} - \bar{Y}_{l\bullet})^2}{n - l}, \\ &= \frac{\text{RSS}}{n - l}.\end{aligned}$$

Remarque : RSS n'est pas divisé par n mais par $n - l$, le nombre de résidus linéairement indépendants.

On dit que la série des résidus a $n - l$ **degrés de liberté**.

► Ecart-type résiduel dans \mathbb{R}

Analyse de la variance

Tester un effet : **comparer les modèles nuls et non-nuls**

Or, la qualité d'ajustement d'un modèle est mesurée par la **somme des carrés des résidus** :

$$\text{RSS} = \sum_{i=1}^I \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet})^2, \quad [\text{modèle non-nul}]$$

$$\text{RSS}_0 = \sum_{i=1}^I \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{\bullet\bullet})^2. \quad [\text{modèle nul}]$$

où $\bar{Y}_{\bullet\bullet} = \frac{n_1}{n} \bar{Y}_{1\bullet} + \frac{n_2}{n} \bar{Y}_{2\bullet} + \dots + \frac{n_I}{n} \bar{Y}_{I\bullet}$.

► Sommes des carrés des résidus dans \mathbb{R}

Évaluation de la qualité d'ajustement par le R^2

On peut évaluer la force d'un effet par le rapport suivant :

$$R^2 = \frac{RSS_0 - RSS}{RSS_0} = \frac{\sum_{i=1}^I n_i (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2}{RSS_0}.$$

En effet,

- $0 \leq R^2 \leq 1$,
- $R^2 = 0$ correspond à 'pas d'effet',
- $R^2 = 1$ correspond à 'effet total'.

► Coefficient R^2 dans R

Test de Fisher d'un effet

La statistique du **test de Fisher** pour un effet groupe sur Y est :

$$F = \frac{(RSS_0 - RSS)/(l - 1)}{RSS/(n - l)}.$$

Degrés de liberté de $RSS_0 - RSS$: $l - 1$

En effet, la somme des variations *inter-groupes* $n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})$ est nulle : seulement $l - 1$ sont linéairement indépendantes.

► Statistique de Fisher dans \mathbb{R}

p-value

Sous H_0 , la statistique F suit la loi de **Fisher** $\mathcal{F}_{l-1, n-l}$ à $l - 1$ et $n - l$ degrés de liberté.

- ▶ Table d'analyse de la variance dans \mathbb{R}

p-value

La 1ère ligne de la **table d'Analyse de la Variance** mesure l'effet groupe et la 2nde mesure l'erreur résiduelle :

- **Df** : d.d.l., respectivement $I - 1$ et $n - I$;
- **Sum Sq** : sommes de carrés, respectivement $RSS_0 - RSS$ et RSS ;
- **Mean Sq** : moyennes de carrés, respectivement $(RSS_0 - RSS)/(I - 1)$ et $RSS/(n - I)$;
- **F value** : Statistique F, le rapport entre les moyennes de carrés;
- **Pr (>F)** : p-value du test de Fisher.

Effet groupe : démarche d'analyse - Résumé

- 1 Modèle d'analyse de la variance à un facteur (I groupes)

Y_{ij} = valeur de Y pour l'individu j du groupe i ,

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}; \varepsilon_{ij} \sim \mathcal{N}(0; \sigma)$$

- 2 H_0 : pas d'effet groupe
- 3 Statistique de test :

$$F = \frac{(\text{RSS}_0 - \text{RSS}) / (I - 1)}{\text{RSS} / (n - I)}.$$

- 4 Loi de F sous H_0 : $\mathcal{F}_{I-1, n-I}$

... on en déduit la p-value.

Plan

- 1 Effet à l'échelle d'une population
- 2 Décider à partir de données
- 3 Effet 'groupe'
 - Comparaison de groupes
 - Analyse de variance à un facteur
 - Estimation des paramètres d'effet
 - Test de Fisher
 - Le cas particulier de la comparaison de 2 groupes
 - Décrire un effet groupe
 - Test avec des données appariées
- 4 Effet linéaire