Démarche statistique
Premiers pas avec R

David Causeur
Agrocampus Ouest
IRMAR CNRS UMR 6625
http://math.agrocampus-ouest.fr/infoglueDeliverLive/membres/david.causeur

Plan

Plan

Plan

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 ($j = 1, \ldots, n_i$) du $i$ème groupe ($i = 1, \ldots, I$) :

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

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 ;
  $\Rightarrow$ les μ_i sont / **paramètres inconnus**.

- ε_{ij} pour les variations aléatoires intra-groupes.
  $\Rightarrow$ σ, l'**écart-type résiduel**, est un autre paramètre inconnu.

Test d'un effet

Test de l'effet de X sur Y :

$$\begin{cases} \mathrm{H}_0 & : \ \mu_1 = \ldots = \mu_I = \mu \ \ \text{(pas d'effet de } X\text{)} \\ \mathrm{H}_1 & : \ \text{Pour au moins un couple } (i, i'), \text{ avec } i \neq i', \ \mu_i \neq \mu_{i'}. \end{cases}.$$

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** :

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

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

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

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 ?

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

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

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

- $\hat{\mu}_i - \mu_i$ 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 R

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

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

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

Écart-type résiduel

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

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

$$\begin{aligned} \hat{\sigma}^2 &= \frac{\sum_{j=1}^{n1}(Y_{1j} - \overline{Y}_{1\bullet})^2 + \ldots + \sum_{j=1}^{nI}(Y_{Ij} - \overline{Y}_{I\bullet})^2}{n - I}, \\ &= \frac{\text{RSS}}{n - I}. \end{aligned}$$

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

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

▶ Ecart-type résiduel dans `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** :

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

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

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

▶ Sommes des carrés des résidus dans `R`

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

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

$$\mathrm{R}^2 \quad = \quad \frac{\mathrm{RSS}_0 - \mathrm{RSS}}{\mathrm{RSS}_0} = \frac{\sum_{i=1}^{I} n_i(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2}{\mathrm{RSS}_0}.$$

En effet,

- $0 \leq \mathrm{R}^2 \leq 1$,
- $\mathrm{R}^2 = 0$ correspond à 'pas d'effet',
- $\mathrm{R}^2 = 1$ 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 :

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

Degrés de liberté de $\mathrm{RSS}_0 - \mathrm{RSS}$: $I - 1$
En effet, la somme des variations *inter-groupes* $n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})$ est nulle : seulement $I - 1$ sont linéairement indépendantes.

▶ Statistique de Fisher dans R

p-value

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

- Table d'analyse de la variance dans 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 $\mathrm{RSS}_0 - \mathrm{RSS}$ et RSS ;
- `Mean Sq` : moyennes de carrés, respectivement $(\mathrm{RSS}_0 - \mathrm{RSS})/(I - 1)$ et $\mathrm{RSS}/(n - I)$;
- `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)

$$\begin{aligned} Y_{ij} &= \text{valeur de } Y \text{ pour l'individu } j \text{ du groupe } i, \\ Y_{ij} &= \mu + \alpha_i + \varepsilon_{ij};\ \varepsilon_{ij} \sim \mathcal{N}(0;\sigma) \end{aligned}$$

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

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

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

... on en déduit la p-value.

Plan