Statistiques et Applications

Analyse de la variance

aurore.lavigne@univ-lille.fr

Situation du problème

- On désire connaître l'effet du sexe sur le salaire :

Situation du problème

- On désire connaître l'effet du sexe sur le salaire :

  => 2 modalités Homme - Femme

  => test de Student de comparaison de moyennes

Situation du problème

- On désire connaître l'effet du sexe sur le salaire :
  => 2 modalités Homme - Femme
  => test de Student de comparaison de moyennes

- On désire connaître l'effet du type de contrat sur le salaire :

Situation du problème

- On désire connaître l'effet du sexe sur le salaire :
  => 2 modalités Homme - Femme
  => test de Student de comparaison de moyennes

- On désire connaître l'effet du type de contrat sur le salaire :
  => 4 modalités CDD - CDI - Apprenti - Occasionnel

Situation du problème

- On désire connaître l'effet du sexe sur le salaire :

  => 2 modalités Homme - Femme

  => test de Student de comparaison de moyennes

- On désire connaître l'effet du type de contrat sur le salaire :

  => 4 modalités CDD - CDI - Apprenti - Occasionnel

  => Analyse de la variance

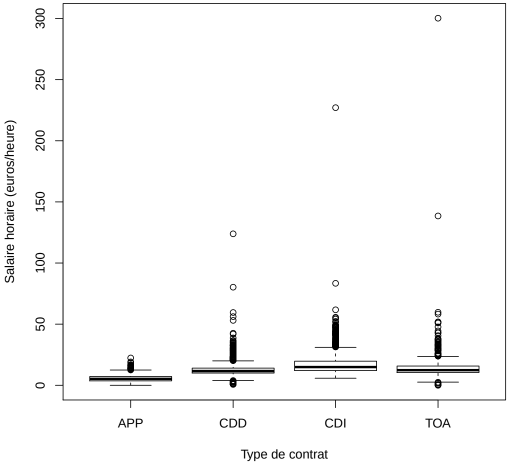


FIGURE – Salaire horaire en fonction du type de contrat. Source : Rééchantillonnage de la base Postes, 2013, INSEE

Analyse de la variance

L'analyse de la variance offre un cadre d'analyse rigoureux pour l'estimation et le test de l'**effet d'une ou plusieurs variables qualitatives sur une variable quantitative**.

VOCABULAIRE :

Les variables qualitatives s'appellent **les facteurs de variabilité** et leurs modalités des **niveaux**. La variable qualitative est la **réponse**.

Plan

Les données et le modèle

On considère **1 facteur** à **k niveaux**.
Pour chaque niveau $l \in \{1, 2, \cdots, k\}$, on dispose d'un échantillon de taille n_i d'observations de la variable quantitative.

$$
\begin{array}{ll}
\text{niveau } 1 & Y_1^1, Y_2^1, \cdots, Y_{n_1}^1 \\
\text{niveau } 2 & Y_1^2, Y_2^2, \cdots, Y_{n_2}^2 \\
\vdots & \vdots \\
\text{niveau } k & Y_1^k, Y_2^k, \cdots, Y_{n_k}^k
\end{array}
$$

Indépendance

On suppose que les **toutes** les variables sont **indépendantes**.

- Les variables d'un même niveau sont indépendantes :
  $\forall l \in \{1, 2, \cdots, k\}$, $\forall i \neq j$, Y_i^l et Y_j^l sont indépendantes.
- Les variables de deux niveaux différents sont indépendantes
  $\forall l \neq m$, $\forall (i, j)$, Y_i^l et Y_j^m sont indépendantes.

Modèle

On suppose de plus que

- toutes les variables suivent une distribution normale
- **l'espérance dépend du niveau** k
- la variance est identique pour toutes les variables

$$Y_i^l \sim \mathcal{N}(\mu_l, \sigma^2), \quad \forall l \in \{1, \cdots, k\}, \quad \forall i \in \{1, \cdots, n_l\}$$

Modèle

De manière équivalente, on pourra écrire que

$$Y_i^l = \mu_l + \epsilon_i^l \quad \text{avec} \quad \epsilon_i^l \sim \mathcal{N}(0, \sigma^2) \text{ et ind.}$$

- μ_l est l'espérance observée pour le niveau l du facteur.

L'anova est un modèle linéaire

En effet on peut réécrire le modèle ci-dessus de la manière suivante :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \text{ avec } \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$$

et

$$\mathbf{X} = \begin{pmatrix}
\overbrace{\begin{matrix} 1 & 0 & \cdots & 0 & \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & 0 & \cdots & 0 & \end{matrix}}^{k+1 \text{ col.}} & \left.\vphantom{\begin{matrix} 1 \\ \vdots \\ 1 \end{matrix}}\right\} n_1 \text{ li.} \\
\begin{matrix} 0 & 1 & \cdots & 0 & \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 0 & 1 & \cdots & 0 & \end{matrix} & \left.\vphantom{\begin{matrix} 1 \\ \vdots \\ 1 \end{matrix}}\right\} n_2 \text{ li.} \\
\begin{matrix} \vdots & \vdots & \vdots & \ldots & \vdots \end{matrix} & \\
\begin{matrix} 0 & 0 & \cdots & 1 & \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \cdots & 1 & \end{matrix} & \left.\vphantom{\begin{matrix} 1 \\ \vdots \\ 1 \end{matrix}}\right\} n_k \text{ li.}
\end{pmatrix} \text{ et } \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix}$$

En conséquence...

... tous les résultats vus sur le modèle linéaire dans le chapitre sur la régression multiple s'appliquent ici, et notamment :

- **l'estimation** et les propriétés des estimateurs
- **les tests sur les paramètres**, ou les comb. lin. de paramètres
- **les tests de comparaison de modèles** en particulier le test de validité globale du modèle.
- **les résidus** leur loi et propriétés.

Plan

Problématique

On cherche à savoir si **le facteur** (la variable qualitative) à un effet sur **la réponse** (la variable quantitative).

=> Le test de validité globale du modèle permet de répondre à cette question.

Les deux hypothèses testées sont

- $\mathcal{H}_0 : \{\mu_1 = \mu_2 = \cdots = \mu_k\}$
- $\mathcal{H}_1 : \{\exists l, m \in \{1, 2, \cdots, k\}$ tels que $\mu_l \neq \mu_m\}$

Décomposition de la variance

Dans le cadre d'analyse de la variance, on a :

$$SCT = \sum_{l=1}^{k}\sum_{i=1}^{n_l}(Y_i^l - \overline{Y})^2 \quad SCM = \sum_{l=1}^{k} n_l(\overline{Y}_l - \overline{Y})^2 \quad SCR = \sum_{l=1}^{k}\sum_{i=1}^{n_l}(Y_i^l - \overline{Y}_l)^2$$

avec $\overline{Y}_l$ la moyenne pour le niveau l : $\overline{Y}_l = \frac{1}{n_l}\sum_{i=1}^{n_l} X_i^l$.

Remarque : L'espace engendré par les colonnes de $\mathbf{X}$ est de dimension k, le degré de liberté associé à SCM est donc $k-1$.

Sous l'hypothèse $\mathcal{H}_0$

$$\frac{SCM/k-1}{SCR/n-k} \sim \mathcal{F}_{k-1,n-k}$$

Table d'analyse de la variance

Source	Somme des carrés	Degrés de liberté	Som. carrés moyens	Statistique F	Proba. crit.
Facteur	SCM	$k-1$	$\frac{SCM}{k-1}$	$\frac{SCM/k-1}{SCR/n-k}$	p_c
Résidu	SCR	$n-k$	$\frac{SCR}{n-k}$		
Total	SCT	$n-1$			

Estimation de l'écart-type σ

$$S = \sqrt{\frac{SCR}{n-k}}$$

est un estimateur de l'écart-type σ.

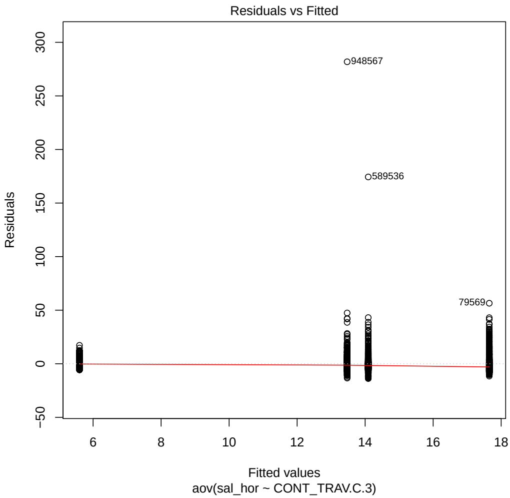
REMARQUES :

- On peut trouver l'estimation de l'écart-type dans la table d'analyse de la variance.
- La quantité $\frac{SCR}{n-k}$ est souvent nommée MSE pour *mean square error* dans les logiciels de statistiques.

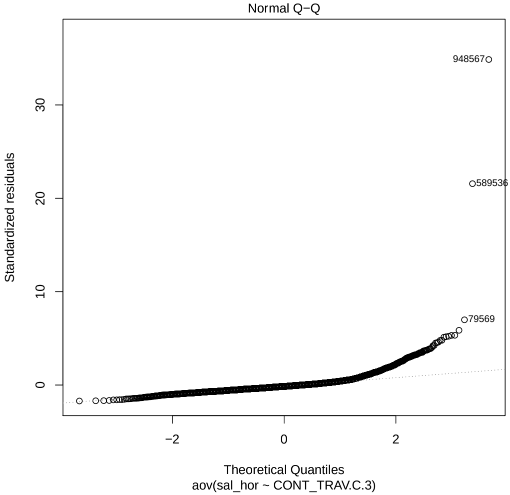
Illustration

```
                Df Sum Sq Mean Sq F value Pr(>F)
CONT_TRAV.C.3    3  77452   25817   394.7 <2e-16 ***
Residuals     3996 261381      65
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Vérification des hypothèses : homoscédasticité



Vérification des hypothèses : normalité



Plan

Ecriture singulière

La plupart des logiciels de statistiques utilisent l'écriture suivante (écriture singulière)

$$Y_i^l = \mu + \alpha_l + \epsilon_i^l \quad \text{avec} \quad \epsilon_i^l \sim \mathcal{N}(0, \sigma^2) \text{ et ind.}$$

- μ est la moyenne générale.
- α_l est l'**effet** du niveau l du facteur

Dans ce cas la matrice $\mathbf{X}$ devient singulière, la première colonne état la somme des k colonnes suivantes.

$$\mathbf{X} = \left(\begin{array}{l} \overbrace{\begin{matrix} 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & 1 & 0 & \cdots & 0 \end{matrix}}^{k+1 \text{ col.}} \left.\vphantom{\begin{matrix}1\\ \vdots\\ 1\end{matrix}}\right\} n_1 \text{ li.} \\ \begin{matrix} 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & 0 & 1 & \cdots & 0 \end{matrix} \left.\vphantom{\begin{matrix}1\\ \vdots\\ 1\end{matrix}}\right\} n_2 \text{ li.} \\ \begin{matrix} \vdots & \vdots & \vdots & \ldots & \vdots \end{matrix} \\ \begin{matrix} 1 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{matrix} \left.\vphantom{\begin{matrix}1\\ \vdots\\ 1\end{matrix}}\right\} n_k \text{ li.} \end{array} \right) \quad \text{et } \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{pmatrix}$$

Identifiabilité

Définition

Soit $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ un modèle statistique. On dit que $\mathcal{P}$ est identifiable si et seulement si

$$P_{\theta_1} = P_{\theta_2} \Longrightarrow \theta_1 = \theta_2, \quad \text{pour tout } \theta_1, \theta_2 \in \Theta.$$

EXEMPLE : Le modèle de Poisson

$$\{\mathcal{P}(\theta) \text{ tq } \theta \in \mathbb{R}^+\}$$

Cas de l'écriture signulière

Modèle

$$Y_i^l = \mu + \alpha_l + \epsilon_i^l \text{ avec } \epsilon_i^l \sim \mathcal{N}(0, \sigma^2), \text{ ind.}$$

Ici, les deux jeux de paramètres $(\mu, \alpha_1, \cdots, \alpha_l, \sigma)$ et $(\mu - 1, \alpha_1 + 1, \cdots, \alpha_l + 1, \sigma)$, pour une valeur fixée de $\mu, \alpha_1, \cdots, \alpha_l$ et σ conduisent à la même probabilité. => Le modèle n'est pas identifiable.

CONSÉQUENCE : On ne peut pas estimer $\mu, \alpha_1, \cdots, \alpha_l$.

Ajout de contraintes d'identifiabilité

Pour rendre le modèle identifiable, on va ajouter une contrainte sur une combinaison linéaire des paramètres $\mu, \alpha_1, \cdots, \alpha_l$. Par exemple :

$$\begin{cases} \sum_{l=1}^{k} n_l \alpha_l = 0 \qquad (1) \\ \alpha_1 = 0 \qquad (2) \end{cases}$$

Avec la contrainte (1), on estime

- μ par $\overline{y}$
- α_l par $\overline{y}_l - \overline{y}$

Avec la contrainte (2), on estime

- μ par $\overline{y}_1$
- α_l par $\overline{y}_l - \overline{y}_1$

ATTENTION :

- Selon les contraintes, les coefficients s'interprètent différemment.
- Si on ne connait pas les contraintes, il ne faut pas chercher à interpréter les tests, et les coefficients.

Illustration avec R

```
  > model.tables(aov_cont,type='effects')
Tables of effects

 CONT_TRAV.C.3
CONT_TRAV.C.3
   APP    CDD    CDI    TOA
-7.104  0.768  4.947  1.388
```

Tests sur les effets

De nombreux logiciels donnent la probabilité critique du test $\mathcal{H}_0 = \{\alpha_l = 0\}$ contre $\mathcal{H}_1 = \{\alpha_l \neq 0\}$.

ATTENTION :
Selon la contrainte utilisée, la signification du test n'est pas la même.

- **Avec (2)** : Le test de $\{\alpha_1 = 0\}$ revient à tester $\{\mu_1 = 0\}$. "La moyenne du groupe 1 est nulle".
  Le test de $\{\alpha_2 = 0\}$ revient à tester $\{\mu_2 = \mu_1\}$. "La moyenne du groupe 2 est égale à la moyenne du groupe 1".
- **Avec (1)** : Le test de $\{\alpha_1 = 0\}$ revient à tester $\{\mu_1 = \mu\}$. "La moyenne du groupe 1 est égale à la moyenne générale".

Illustration

```
> model.tables(aov_cont,type='means')
Tables of means
Grand mean

12.70279

 CONT_TRAV.C.3
CONT_TRAV.C.3
   APP    CDD    CDI    TOA
 5.599 13.471 17.650 14.091
```

Plan

Contraste

Définition

On appelle contraste L des k moyennes $\mu_1, \mu_2, \cdots, \mu_k$ la somme

$$L = \sum_{l=1}^{k} l_l \mu_l \text{ telle que } \sum_{l=1}^{k} l_l = 0.$$

Exemples :

- $\mu_1 - \mu_2$: pour comparer μ_1 à μ_2
- $\mu_1 - 2\mu_2 + \mu_3$: pour comparer μ_2 à la moyenne de μ_1 et μ_3.

Estimation

Un estimateur sans biais de L est

$$\hat{L} = \sum_{l=1}^{l} l_l \hat{\mu}_l = \sum_{l=1}^{l} l_l \overline{X}_l$$

Propriétés

On a

- $(\hat{L}) = L$
- $V(\hat{L}) = \sigma^2 \sum_{l=1}^{k} \frac{l_l^2}{n_l}$
- $$\frac{\hat{L} - L}{S\sqrt{\sum_{l=1}^{k} \frac{l_l^2}{n_l}}} \sim \mathcal{S}_{n-k}$$

Tests sur les contrastes

Tests *a priori*

On sait *a priori* à quelle question doit répondre notre analyse. On définit le contraste en fonction de la problématique et on test $\mathcal{H}_0 = \{L = 0\}$ contre $\mathcal{H}_1 = \{L \neq 0\}$.

- Avantages : on réalise peu de tests
- Inconvénients : il faut à l'avance savoir ce que l'on veut tester

Comparaisons multiples *a posteriori*

On ne sait pas *a priori* ce que l'on cherche, on se trouve dans une démarche exploratoire. On teste tous les contrastes $\mu_l - \mu_{l'}$.

- Avantages : on n'a pas besoin d'avoir une question par avance.
- Inconvénients : tests multiples, on réalise $\frac{k(k-1)}{2}$ tests.

Tests multiples

Soit une famille de m hypothèses de tests $\mathcal{H}_{0i}$ contre $\mathcal{H}_{1i}$, pour $i \in \{1, 2, \cdots, m\}$.

Definition

On appelle $FWER$ le *family wise error rate*, la probabilité de rejeter à tort au moins 1 fois une hypothèse $\mathcal{H}_{0i}$ sur les m tests réalisés.

Propriété

Si les m tests sont indépendants et tous de niveau α alors

$$FWER = 1 - (1 - \alpha)^m$$

$=>$ Démonstration

m	1	5	10	20	100
$FWER$	0.05	0.22	0.40	0.64	0.99

CONSÉQUENCE : On ne contrôle plus le risque de première espèce.

Méthode de Bonferroni

On diminue le risque de première espèce α. On prend $\alpha' = \frac{\alpha}{m}$.

- Avantage : on diminue la probabilité de réaliser au moins une erreur de première espèce sur les m tests.
- Inconvénient : on diminue aussi la puissance du test. On aura des difficultés à repérer les groupes différents.

Etendue Studentisée

Définition

On suppose que $Z_1, Z_2, \cdots, Z_m \sim \mathcal{N}(0,1)$ sont m variables normales standardisées indépendantes. On suppose que $U \sim \chi^2_\nu$ est aussi indépendante des Z_i.
L'**étendue Studentisé** est la variable aléatoire :

$$Q_{m,\nu} = \frac{\max_i Z_i - \min_i Z_i}{\sqrt{U/\nu}}$$

Application au cas des comparaisons multiples

On suppose que $n_1 = n_2 = \cdots = n_k = r$.
Nous avons vu que

- $\frac{\overline{X}_l - \mu_l}{\sigma/\sqrt{r}} \sim \mathcal{N}(0,1)$,
- $\overline{X}_1, \overline{X}_2, \cdots, \overline{X}_k$ sont indépendantes.
- $\frac{(n-k)S^2}{\sigma^2} \sim \chi^2_{n-k}$

On en déduit donc que

$$\frac{(\max_l \overline{X}_l - \min_l \overline{X}_l) - (\mu_M - \mu_m)}{S/\sqrt{r}} \sim Q_{k,n-k}$$

avec μ_M (resp. μ_m) l'espérance du groupe M tel que $\mu_M = \max \overline{X}_l$ (resp. m tel que $\mu_m = \min \overline{X}_l$).

Procédure de comparaisons multiples de Tukey

On considère les $m = k(k-1)/2$ hypothèses $\mathcal{H}_{0ll'} = \{\mu_l = \mu_{l'}\}$,
Pour les $k(k-1)/2$ contrastes linéaires $\mu_l - \mu_{l'}$.

1. Calculer la différence
$$|\overline{X}_l - \overline{X}_{l'}|$$
2. Rejeter les hypothèses $\mathcal{H}_{0ll'} = \{\mu_l = \mu_{l'}\}$ si
$$|\overline{X}_l - \overline{X}_{l'}| > R_{crit}$$
avec
$$R_{crit} = Q_{k,n-k,1-\alpha} S/\sqrt{r}$$

Cette procédure permet de contrôler le $FWER$.

Justification de la procédure de Tukey

Si toutes les hypothèses $\mathcal{H}_{0ll'}$ sont vérifiées simultanément, alors,

$$\frac{(\max_l \overline{X}_l - \min_l \overline{X}_l)}{S/\sqrt{r}} \sim Q_{k,n-k}$$

$=>$ Démonstration

Justification de la procédure de Tukey

Si toutes les hypothèses $\mathcal{H}_{0ll'}$ sont vérifiées simultanément, alors,

$$\frac{(\max_l \overline{X}_l - \min_l \overline{X}_l)}{S/\sqrt{r}} \sim Q_{k,n-k}$$

=> Démonstration

La procédure permet de contrôler le $FWER$.

$$\begin{aligned} FWER &= \text{(Au moins une des hypothèses } \mathcal{H}_{0ll'} \text{ est rejetée à tort.)} \\ &= \alpha \end{aligned}$$

=> Démonstration.

Justification de la procédure de Tukey

Lemme

Soient $Z_l = \overline{X}_l - \mu_l$, pour $l \in \{1, 2, \cdots, k\}$ telles que

$$((\max_l \overline{Z}_l - \min_l \overline{Z}_l) < R_{crit}) = 1 - \alpha,$$

alors,

$$(|\overline{Z}_l - \overline{Z}_{l'}| < R_{crit},\ \text{pour tout } l, l') = 1 - \alpha.$$

Intervalle de confiance simultané

Définition

L'intervalle de confiance simultané pour les paramètres $\mu_1, \mu_2, \cdots, \mu_k$ est l'ensemble des points $\mu_{10}, \mu_{20}, \cdots, \mu_{k0}$, tels qu'aucune des $k(k-1)/2$ hypothèses de test $\mathcal{H}_{0ll'} = \{\mu_{l0} - \mu_{l'0}\}$ ne soit rejetée avec la procédure de test.

Probabilité critique ajustée

Définition

La probabilité critique ajustée du test $\mathcal{H}_{0ll'} = \{\mu_l = \mu_{l'}\}$ contre $\mathcal{H}_{1ll'} = \{\mu_l \neq \mu_{l'}\}$ est la plus petite valeur du risque de première espèce α telle que $\mathcal{H}_{0ll'}$ est rejetée par la procédure de test.

La région de rejet dépend de α

p_{adj} est telle que $R_{crit}(p_{adj}) = |\overline{x}_l - \overline{x}_{l'}|$.

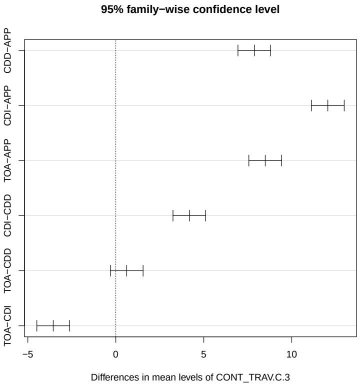
Illustration

```
> TukeyHSD(aov_cont)
Tukey multiple comparisons of means
  95% family-wise confidence level

Fit: aov(formula = sal_hor ~ CONT_TRAV.C.3, data = subdata)

$CONT_TRAV.C.3
            diff        lwr       upr     p adj
CDD-APP  7.87187  6.9422807  8.801459 0.0000000
CDI-APP 12.05058 11.1209907 12.980169 0.0000000
TOA-APP  8.49198  7.5623907  9.421569 0.0000000
CDI-CDD  4.17871  3.2491207  5.108299 0.0000000
TOA-CDD  0.62011 -0.3094793  1.549699 0.3161723
TOA-CDI -3.55860 -4.4881893 -2.629011 0.0000000
```

Illustration



Plan du cours