Analyse de la variance à un facteur

Myriam Maumy-Bertrand [1] & Marie Chion[1]

[1]IRMA, Université de Strasbourg
Strasbourg, France

Master 1re Année
2019-2020

Sommaire

Sommaire

Sommaire

Références

Ce cours s'appuie essentiellement sur

1. le livre David C. Howell, **Méthodes statistiques en sciences humaines** traduit de la sixième édition américaine aux éditions de Boeck, 2008.
2. le livre de Pierre Dagnelie, **Statistique théorique et appliquée**, Tome 2, aux éditions de Boeck, 1998.
3. le livre de Hardeo Sahai et Mohammed I. Ageel, **The Analysis of Variance : Fixed, Random and Mixed Models**, aux éditions Birkhäuser, 2000.

Sommaire

Objectif

Dans ce chapitre, nous allons étudier un test statistique (nous renvoyons à un cours sur les tests pour toutes les définitions sur ce sujet) permettant de comparer les moyennes de plusieurs variables aléatoires indépendantes gaussiennes de même variance.

L'analyse de la variance est l'une des procédures les plus utilisées dans les applications de la statistique ainsi que dans les méthodes d'analyse de données.

Exemple : D'après le livre de William P. Gardiner, Statistical Analysis Methods for Chemists

Une étude de reproductibilité a été menée pour étudier les performances de trois laboratoires relativement à la détermination de la quantité de sodium de lasalocide dans de la nourriture pour de la volaille.

Une portion de nourriture contenant la dose nominale de 85 mg kg^{-1} de sodium de lasalocide a été envoyée à chacun des laboratoires à qui il a été demandé de procéder à 10 réplications de l'analyse.

Les mesures de sodium de lasalocide obtenues sont exprimées en mg kg^{-1}. Elles ont été reproduites sur le transparent suivant.

Exemple : D'après le livre de William P. Gardiner.

	Laboratoire		
	A	*B*	*C*
1	87	88	85
2	88	93	84
3	84	88	79
4	84	89	86
5	87	85	81
6	81	87	86
7	86	86	88
8	84	89	83
9	88	88	83
10	86	93	83

TABLE – Source : Analytical Methods Committee, *Analyst*, 1995.

Remarque

Cette écriture du tableau est dite « désempilée ». Nous pouvons l'écrire sous forme standard (« empilée »), c'est-à-dire avec deux colonnes, une pour le laboratoire et une pour la valeur de sodium de lasalocide mesurée, et trente lignes, une pour chacune des observations réalisées.

Tableau empilé de l'exemple des laboratoires

Essai	Laboratoire	Lasalocide
1	Laboratoire *A*	87
2	Laboratoire *A*	88
3	Laboratoire *A*	84
4	Laboratoire *A*	84
5	Laboratoire *A*	87
6	Laboratoire *A*	81
7	Laboratoire *A*	86
8	Laboratoire *A*	84
9	Laboratoire *A*	88
10	Laboratoire *A*	86

Suite du tableau précédent

Essai	Laboratoire	Lasalocide
11	Laboratoire *B*	88
12	Laboratoire *B*	93
13	Laboratoire *B*	88
14	Laboratoire *B*	89
15	Laboratoire *B*	85
16	Laboratoire *B*	87
17	Laboratoire *B*	86
18	Laboratoire *B*	89
19	Laboratoire *B*	88
20	Laboratoire *B*	93

Suite du tableau précédent

Essai	Laboratoire	Lasalocide
21	Laboratoire C	85
22	Laboratoire C	84
23	Laboratoire C	79
24	Laboratoire C	86
25	Laboratoire C	81
26	Laboratoire C	86
27	Laboratoire C	88
28	Laboratoire C	83
29	Laboratoire C	83
30	Laboratoire C	83

Remarque

Dans la plupart des logiciels, c'est sous cette forme que sont saisies et traitées les données. Dans les deux tableaux, nous avons omis les unités de la mesure réalisée et ceci pour abréger l'écriture. Mais en principe cela doit être indiqué entre parenthèses à côté de la mesure.

Remarque

Il va de soi que lorsque vous rentrerez des données sous un logiciel, vous n'indiquerez pas le mot « Laboratoire » à côté des nombres (A, B, C). Il est juste là pour vous faciliter la compréhension du tableau.

Définitions

Sur **chaque essai**, nous observons **deux variables**.

1. Le laboratoire. Il est totalement contrôlé. La variable « Laboratoire » est considérée comme qualitative avec trois modalités bien déterminées. Nous l'appelons **le facteur (factor)**. Ici le facteur « Laboratoire » est à **effets fixes (fixed effects)**.
2. La quantité de Lasalocide. La variable « Lasalocide » est considérée comme quantitative comme généralement tous les résultats obtenue par une mesure. Nous l'appelons **la réponse (response)**.

Notations

La variable mesurée dans un tel schéma expérimental sera notée Y.

Pour les observations nous utilisons deux indices :

- le premier indice indique le numéro du groupe dans la population (« Laboratoire »),
- le second indice indique le numéro de l'observation dans l'échantillon (« Essai »).

Signification des indices

Pour le premier indice, nous utilisons i (ou encore i', i'', i_1, i_2).
Pour le second indice, nous utilisons j (ou encore j', j'', j_1, j_2).

Notation

Ainsi les observations sont en général notées par :

$$y_{ij}, \quad i = 1, \ldots, I \quad j = 1, \ldots, J(i).$$

Définition

Lorsque *les échantillons sont de même taille, à savoir $J(i) = J$ et ce quelque soit i*, nous disons que l'expérience est **équilibrée**.

Remarque

Si *les tailles des échantillons sont différentes*, alors elles sont notées par :

$$n_i, \quad \text{où} \quad i = 1, \ldots, I.$$

Mais ce plan expérimental est à éviter parce que les différences qu'il est alors possible de détecter sont supérieures à celles du schéma équilibré.

Définitions

En se plaçant dans le **cas équilibré** nous notons les **moyennes (means)** de chaque échantillon par :

$$\overline{y}_i = \frac{1}{J} \sum_{j=1}^{J} y_{ij}, \quad i = 1, \ldots, I,$$

et les **variances (variances)** de chaque échantillon par :

$$s_i^2(y) = \frac{1}{J} \sum_{j=1}^{J} (y_{ij} - \overline{y}_i)^2, \quad i = 1, \ldots, I.$$

Remarque

Cette dernière formule exprime la variance non corrigée. Très souvent, dans les ouvrages ou les logiciels, c'est la variance corrigée qui est utilisée : au lieu d'être divisée par J, la somme est divisée par $J - 1$.

Retour à l'exemple

Après calculs avec le logiciel **R**, nous avons :

$$\overline{y}_1 = 85,500 \quad \overline{y}_2 = 88,600$$
$$\overline{y}_3 = 83,800.$$

et

$$s_{1,c}(y) = 2,224 \quad s_{2,c}(y) = 2,633$$
$$s_{3,c}(y) = 2,616.$$

Le nombre total d'observations est égal à :

$$n = IJ = 3 \times 10 = 30.$$

Conditions fondamentales de l'ANOVA

Les résidus $\{\widehat{e}_{ij}\}$ sont associés, sans en être des réalisations, aux variables erreurs $\{\varepsilon_{ij}\}$ qui sont inobservables et satisfont aux 3 conditions suivantes :

1. Elles sont **indépendantes (independent)**.
2. Elles ont **même variance** σ^2 inconnue. C'est la condition d' **homogénéité (homogeneity)** ou d' **homoscédasticité (homoscedasticity)**.
3. Elles sont de **loi gaussienne (normal distribution)**.

Remarque

Par conséquent ces trois conditions se transfèrent sur les variables aléatoires $\{Y_{ij}\}$.

Modèle statistique

Nous pouvons donc écrire le modèle :

$$\mathcal{L}(Y_{ij}) = \mathcal{N}(\mu_i \,;\, \sigma^2), \quad i = 1, \ldots, I, \quad j = 1, \ldots, J.$$

Ainsi nous constatons que, si les lois $\mathcal{L}(Y_{ij})$ sont différentes, elles ne peuvent différer que par leur moyenne théorique. Il y a donc un simple décalage entre elles.

Remarque

Parfois, le modèle statistique est écrit de la façon suivante :

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

où $\sum_{i=1}^{I} \alpha_i = 0$ et $\mathcal{L}(\varepsilon_{ij}) = \mathcal{N}(0; \sigma^2), \quad i = 1, \ldots, I, \quad j = 1, \ldots, J.$

Nous avons donc la correspondance suivante :

$$\mu_i = \mu + \alpha_i \quad i = 1, \ldots, I.$$

Les deux modèles sont donc statistiquement équivalents.

Mise en place du test de comparaison des moyennes

Nous nous proposons de tester l'hypothèse nulle

$$(\mathcal{H}_0) : \mu_1 = \mu_2 = \cdots = \mu_I$$

contre l'hypothèse alternative

$$(\mathcal{H}_1) : \text{Les moyennes } \mu_i \text{ ne sont pas toutes égales.}$$

La méthode statistique qui permet d'effectuer ce test est appelée l'**analyse de la variance à un facteur (one way analysis of variance)**.

Sommaire

Deux propriétés fondamentales

Le test est fondé sur deux propriétés des moyennes et des variances.

Première propriété

La moyenne de toutes les observations est la moyenne des moyennes de chaque échantillon. Ceci s'écrit :

$$\overline{y} = \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{I} y_{ij} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij} = \frac{1}{I}\sum_{i=1}^{I}\overline{y}_{i}.$$

Retour à l'exemple

Pour cet exemple, nous constatons cette propriété. En effet, nous avons avec le logiciel **R** :

$$\begin{aligned}\overline{y} &= \frac{1}{30} \times 2579 \\ &= \frac{1}{3}(85,500 + 88,600 + 83,800) \\ &= \frac{1}{3} \times 257,900 \\ &= 85,967,\end{aligned}$$

puisque $n = 30 = I \times J = 3 \times 10$.

Deuxième propriété

La variance de toutes les observations est la somme de la variance des moyennes et de la moyenne des variances. Ceci s'écrit :

$$s^2(y) = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij}-\overline{y})^2 = \frac{1}{I}\sum_{i=1}^{I}(\overline{y}_i-\overline{y})^2 + \frac{1}{I}\sum_{i=1}^{I}s_i^2(y). \quad (1)$$

Retour à l'exemple

Un calcul « à la main » avec **R** donne :

$$s^2(y) = 9,566.$$

D'autre part, nous constatons que la variance des moyennes est égale à :

$$\begin{aligned}\frac{1}{I}\sum_{i=1}^{I}(\overline{y}_i-\overline{y})^2 &= \frac{1}{3}\Big((85,500-85,967)^2+(88,600-\\ &\quad 85,967)^2+(83,800-85,967)^2\Big)\\ &= 3,949.\end{aligned}$$

Suite de l'exemple

Nous constatons également que la moyenne des variances est égale à :

$$\frac{1}{I}\sum_{i=1}^{I}s_i^2(y) = \frac{1}{3}(4,450+6,240+6,160) = 5,617.$$

En faisant la somme des deux derniers résultats, nous retrouvons bien la valeur de 9,566 que nous avons obtenue par le calcul simple. Donc la relation (1) est bien vérifiée.

Résultat fondamental de l'ANOVA

En multipliant les deux membres par n de l'équation (1), nous obtenons :

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij}-\overline{y})^2 = J\sum_{i=1}^{I}(\overline{y}_i-\overline{y})^2 + \sum_{i=1}^{I}\left(\sum_{j=1}^{J}(y_{ij}-\overline{y}_i)^2\right)$$

ou encore ce qui s'écrit :

$$sc_{Tot} = sc_F + sc_R. \quad (2)$$

Retour à l'exemple

Avec le logiciel **R**, nous avons d'une part

$$sc_{Tot} = 286,967$$

et d'autre part

$$sc_F = 118,467 \quad \text{et} \quad sc_R = 168,500.$$

Donc lorsque nous faisons la somme des deux derniers résulats nous retrouvons bien la valeur du premier résultat. Donc la relation (2) est bien vérifiée.

Définition

Nous appelons **variation totale (total variation)** le terme :

$$sc_{Tot} = \sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - \overline{y})^2.$$

Elle indique la dispersion des données autour de la moyenne générale.

Définition

Nous appelons **variation due au facteur (variation between)** le terme :

$$sc_F = J\sum_{i=1}^{I}(\overline{y}_i - \overline{y})^2.$$

Elle indique la dispersion des moyennes autour de la moyenne générale.

Définition

Nous appelons **variation résiduelle (variation within)** le terme :

$$sc_R = \sum_{i=1}^{I}\left(\sum_{j=1}^{J}(y_{ij} - \overline{y}_i)^2\right).$$

Elle indique la dispersion des données à l'intérieur de chaque échantillon autour de sa moyenne.

Principe du test :

Si l'hypothèse nulle $(\mathcal{H}_0)$ est vraie alors la quantité SC_F doit être petite par rapport à la quantité SC_R.

Par contre, si l'hypothèse alternative $(\mathcal{H}_1)$ est vraie alors la quantité SC_F doit être grande par rapport à la quantité SC_R.

Pour comparer ces quantités, R. A. Fisher, après les avoir « corrigées » par leurs degrés de liberté (*ddl*), a considéré leur rapport.

Définition

Nous appelons **carré moyen associé au facteur** le terme

$$CM_F = \frac{SC_F}{I-1}$$

et **carré moyen résiduel** le terme

$$CM_R = \frac{SC_R}{n-I}.$$

Propriété

Le **carré moyen résiduel** est un estimateur sans biais de la variance des erreurs σ^2.

C'est pourquoi il est souvent également appelé **variance résiduelle** et presque systématiquement noté S_R^2 lorsqu'il sert à estimer la variance des erreurs.

Sa valeur observée sur l'échantillon est ainsi notée cm_R ou s_R^2.

Propriété

Si les **trois conditions** sont satisfaites et si l'hypothèse nulle $(\mathcal{H}_0)$ est vraie alors

$$F_{obs} = \frac{cm_F}{cm_R}$$

est une réalisation d'une variable aléatoire F qui suit une loi de Fisher à $I-1$ degrés de liberté au numérateur et $n-I$ degrés de liberté au dénominateur. Cette loi est notée $\mathcal{F}_{I-1,n-I}$.

Décision

Pour un seuil donné α (=5%=0,05 en général), les tables de Fisher nous fournissent une valeur critique c telle que $\mathbb{P}_{(\mathcal{H}_0)}[F \leqslant c] = 1 - \alpha$. Alors nous décidons :

$$\begin{cases} \text{si } F_{obs} < c & (\mathcal{H}_0) \text{ est vraie ,} \\ \text{si } c \leqslant F_{obs} & (\mathcal{H}_1) \text{ est vraie .} \end{cases}$$

Tableau de l'ANOVA

L'ensemble de la procédure est résumé par un tableau, appelé **tableau de l'analyse de la variance (analysis of variance table)**, du type suivant :

Variation	SC	ddl	CM	F_{obs}	F_c
Due au facteur	sc_F	$I-1$	cm_F	$\frac{cm_F}{cm_R}$	c
Résiduelle	sc_R	$n-I$	cm_R		
Totale	sc_{Tot}	$n-1$			

Retour à l'exemple

Pour les données de l'exemple des laboratoires, le tableau de l'analyse de la variance s'écrit :

Variation	SC	ddl	CM	F_{obs}	F_c
Due au facteur	118,467	2	59,233	9,49	3,35
Résiduelle	168,500	27	6,241		
Totale	286,967	29			

Conclusion

Pour un seuil $\alpha = 5\%$, les tables de Fisher nous fournissent la valeur critique $c = 3,35$. Le test est significatif puisque $9,49 \geqslant 3,35$. Nous décidons donc de rejeter l'hypothèse nulle nulle $(\mathcal{H}_0)$ est vraie et de décider que l'hypothèse alternative $(\mathcal{H}_1)$ est vraie : il y a une différence entre les moyennes théoriques des quantités de lasalocide entre les laboratoires. Le risque associé à cette décision est un risque de première espèce qui vaut $\alpha = 5\%$.

Nous en concluons que la quantité de lasalocide mesurée varie significativement d'un laboratoire à l'autre.

Remarque

- Nous avons décidé que les moyennes théoriques sont différentes dans leur ensemble, mais nous aurions très bien pu trouver le contraire.
- Comme nous avons décidé que **les moyennes théoriques** sont **différentes** dans leur ensemble que le facteur étudié est **à effets fixes** et qu'il a **plus de trois modalités**, nous pourrions essayer de déterminer là où résident les différences avec un des tests de **comparaisons multiples** détaillés à la Section 4.

Sommaire

Vérification des trois conditions

Nous étudions les possibilités d'évaluer la validité des **trois conditions** que nous avons supposées satisfaites.

Condition d'indépendance

Il n'existe pas, dans un contexte général, **de test statistique simple permettant d'étudier l'indépendance.**

Ce sont les conditions de l'expérience qui nous permettront d'affirmer que nous sommes dans le cas de l'indépendance.

Condition de normalité

Nous ne pouvons pas, en général, la tester pour chaque échantillon. En effet le nombre d'observations est souvent très limité pour chaque échantillon.

Nous allons donc la tester sur l'ensemble des données.

Remarque

Remarquons que si les conditions sont satisfaites et si nous notons :

$$\mathcal{E}_{ij} = Y_{ij} - \mu_i,$$

alors

$$\mathcal{L}(\mathcal{E}_{ij}) = \mathcal{N}(0\ ;\ \sigma^2),$$

alors c'est la même loi pour l'ensemble des unités.

Les moyennes μ_i étant inconnues, nous les estimons par les estimateurs de la moyenne : les $\overline{Y}_i$ où ils sont définis par

$$\overline{Y}_i = \frac{1}{J}\sum_{j=1}^{J} Y_{ij}.$$

Suite de la remarque

Nous obtenons alors les estimations $\overline{y}_i$. Les quantités obtenues s'appellent les **résidus (residuals)** et sont notées $\widehat{e}_{ij}$. Les résidus s'expriment par :

$$\widehat{e}_{ij} = y_{ij} - \overline{y}_i, \quad i = 1, \ldots, I, \quad j = 1, \ldots, J.$$

Les résidus peuvent s'interpréter comme des estimations des erreurs de mesure.

Tests utilisés pour tester la normalité

Nous pouvons alors tester la normalité, avec le **test de Shapiro-Wilk** ou avec le **test de Shapiro-Francia** sur l'ensemble des résidus.

Hypothèses

Nous notons $\widehat{\mathcal{E}}_{ij}$ la variable aléatoire dont le résidu $\widehat{e}_{ij}$ est la réalisation.
L'hypothèse nulle

$$(\mathcal{H}_0) : \mathcal{L}(\widehat{\mathcal{E}}_{ij}) = \mathcal{N}$$

contre l'hypothèse alternative

$$(\mathcal{H}_1) : \mathcal{L}(\widehat{\mathcal{E}}_{ij}) \neq \mathcal{N}.$$

Décision pour le test de Shapiro-Francia

Pour un seuil donné α ($=$ 5% en général), les tables de Shapiro-Francia nous fournissent une valeur critique c telle que $\mathbb{P}_{(\mathcal{H}_0)}[R \leqslant c] = \alpha$. Alors nous décidons :

$$\begin{cases} \text{si } r_{obs} \leqslant c & (\mathcal{H}_1) \text{ est vraie,} \\ \text{si } c < r_{obs} & (\mathcal{H}_0) \text{ est vraie.} \end{cases}$$

Remarque : Dans le cadre de ce cours, la statistique de Shapiro-Francia ne sera jamais calculée. L'utilisateur connaîtra toujours la valeur r_{obs}.

Retour à l'exemple : le test de Shapiro-Francia

Pour un seuil $\alpha = 5\%$, les tables de Shapiro-Francia (qui sont à télécharger sur le site) nous fournissent, avec $n = 30$, la valeur critique $c = 0,9651$. Mais nous avons $r_{obs} = 0,9803$. Comme $c < r_{obs}$, l'hypothèse nulle $(\mathcal{H}_0)$ ne peut être rejetée, c'est-à-dire que **nous décidons que l'hypothèse de normalité est satisfaite.**

Retour à l'exemple : le test de Shapiro-Wilk

Avec le logiciel **R**, nous avons

```
> shapiro.test(residuals(modele))
Shapiro-Wilk normality test
data: residuals(modele)
W = 0.9737, p-value = 0.6431
```

Comme la $p-$valeur est supérieure à 0,05, l'hypothèse nulle $(\mathcal{H}_0)$ est vraie, c'est-à-dire que **nous décidons que l'hypothèse de normalité est satisfaite.**

Condition d'homogénéité

Plusieurs tests permettent de tester l'égalité de plusieurs variances. Parmi ceux-ci, le test le plus utilisé est le **test de Bartlett** dont le protocole est le suivant :

Hypothèses

L'hypothèse nulle

$$(\mathcal{H}_0) : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_I^2$$

contre l'hypothèse alternative

$$(\mathcal{H}_1) : \text{Les variances } \sigma_i^2 \text{ ne sont pas toutes égales.}$$

Statistique

$$B_{obs} = \frac{1}{C_1}\left[(n-I)\ln(s_R^2) - \sum_{i=1}^{I}(n_i-1)\ln(s_{c,i}^2)\right] \qquad (3)$$

où

- la quantité C_1 est définie par :

$$C_1 = 1 + \frac{1}{3(I-1)}\left(\left(\sum_{i=1}^{I}\frac{1}{n_i-1}\right) - \frac{1}{n-I}\right),$$

- s_R^2 la variance résiduelle, $s_{c,i}^2$ la variance corrigée des observations de l'échantillon d'ordre i, $(i = 1, \ldots, I)$.

Propriété

Sous l'hypothèse nulle $(\mathcal{H}_0)$ le nombre B_{obs} défini par (3) est la réalisation d'une variable aléatoire B qui suit asymptotiquement une loi du khi-deux à $I-1$ degrés de liberté.

En pratique, nous pouvons l'appliquer lorsque les effectifs n_i des I échantillons sont tous au moins égaux à 3.
Ce test requiert la normalité des erreurs.

Décision

Pour un seuil donné α ($= 5\%$ en général), les tables du khi-deux nous fournissent une valeur critique c telle que $\mathbb{P}_{(\mathcal{H}_0)}[B \leqslant c] = 1 - \alpha$. Alors nous décidons :

$$\begin{cases} \text{si } c \leqslant B_{obs} & (\mathcal{H}_1) \text{ est vraie,} \\ \text{si } B_{obs} < c & (\mathcal{H}_0) \text{ est vraie.} \end{cases}$$

Retour à l'exemple

En se souvenant que **les** n_i **sont tous égaux**, nous avons, avec le logiciel **R** :

$$B_{obs} = 0,3024.$$

Pour un seuil $\alpha = 5\%$ la valeur critique d'un khi-deux à 2 degrés de liberté, est $c = 5,991$.

Comme $B_{obs} < c$, nous décidons que l'hypothèse nulle $(\mathcal{H}_0)$ ne peut être rejetée, c'est-à-dire que l'hypothèse d'homogénéité des variances est vérifiée.

Sommaire

Objectif

Lorsque pour la comparaison des moyennes théoriques la décision est « l'hypothèse alternative $(\mathcal{H}_1)$ est vraie », pour analyser les différences nous procédons à des tests qui vont répondre à la question suivante :

- D'où vient la différence ?
- Quelles moyennes sont différentes ?

Ces tests qui vont répondre à cette question sont les tests de comparaisons multiples, des adaptations du test de Student.

Comparaison a priori et a posteriori

Les méthodes de comparaison de moyennes à utiliser sont classées en comparaison *a priori* et *a posteriori*

- *A priori*
  Avant de faire l'expérience, l'expérimentateur connaît la liste des hypothèses qu'il veut tester.
  Exemple : montrer que les deux premiers laboratoires sont différents des deux autres.
  Méthodes :
  - Méthode de Bonferroni,
  - Méthode des contrastes linéaires.

Modélisation statistique
Tableau de l'analyse de la variance
Vérification des trois conditions
Comparaisons multiples
Un exemple entièrement traité
Méthode de Bonferroni
Méthode des contrastes linéaires
Méthodes basées sur la statistique de rang studentisée
Méthode de Newman Keuls
Méthode de Tukey

Comparaison a priori et a posteriori

- *A posteriori*
  Après l'expérience, l'expérimentateur regarde les résultats et oriente ses tests en fonction de ce qu'il observe dans les données.
  Exemple : prendre la plus grande et la plus petite moyenne et tester si elles sont vraiment différentes.
  Méthodes :
  - Méthode basée sur la statistique de rang studentisée,
  - Méthode de Newman Keuls,
  - Méthode de Tukey HSD

Myriam Maumy-Bertrand & Marie Chion — Analyse de la variance à un facteur

Modélisation statistique
Tableau de l'analyse de la variance
Vérification des trois conditions
Comparaisons multiples
Un exemple entièrement traité
Méthode de Bonferroni
Méthode des contrastes linéaires
Méthodes basées sur la statistique de rang studentisée
Méthode de Newman Keuls
Méthode de Tukey

Correction de Bonferroni

Idée

- Se fixer la liste des c comparaisons à faire et un taux global d'erreur de type I : α.
- Faire chaque comparaison à un seuil $\alpha' = \alpha/c$.

Bonferroni a montré que cette procédure garantit un taux d'erreur global plus faible que α.

k	c	α	α'	P
2	1	$0,05$	$0,0500$	$0,0500$
4	6	$0,05$	$0,0083$	$0,0490$
6	15	$0,05$	$0,0033$	$0,0489$
8	28	$0,05$	$0,0018$	$0,0488$

Myriam Maumy-Bertrand & Marie Chion — Analyse de la variance à un facteur

Modélisation statistique
Tableau de l'analyse de la variance
Vérification des trois conditions
Comparaisons multiples
Un exemple entièrement traité
Méthode de Bonferroni
Méthode des contrastes linéaires
Méthodes basées sur la statistique de rang studentisée
Méthode de Newman Keuls
Méthode de Tukey

Test de Bonferroni

Objectif : Comparer deux à deux toutes les moyennes possibles des I groupes

1. Calculer le nombre de comparaisons : $n_c = (I \times (I-1))/2$
2. Erreur de type I globale : $\alpha = 5\% = 0,05$
3. Erreur pour chaque test : $\alpha' = 0,05/n_c$
4. Hypothèses : $(\mathcal{H}_0) : \mu_i = \mu_j$ contre $(\mathcal{H}_1) : \mu_i \neq \mu_j$
5. Statistique de test : $t_{obs} = \dfrac{\overline{y}_i - \overline{y}_j}{\sqrt{s_R^2\left(\dfrac{1}{n_i} + \dfrac{1}{n_j}\right)}}$
6. Règle de décision : Accepter $(\mathcal{H}_0)$ si $|t_{obs}| < t_{n-I;1-(\alpha'/2)}$

Myriam Maumy-Bertrand & Marie Chion — Analyse de la variance à un facteur

Retour à l'exemple

Objectif : Illustrons la procédure précédente en comparant le laboratoire 2 et le laboratoire 3.

1. Nombre de comparaisons : $n_c = (3 \times 2)/2 = 3$
2. Erreur de type I globale : $\alpha = 5\% = 0,05$
3. Erreur pour ce test : $\alpha' = 0,05/3 \simeq 0,01667$
4. Hypothèses : $(\mathcal{H}_0) : \mu_2 = \mu_3$ contre $(\mathcal{H}_1) : \mu_2 \neq \mu_3$
5. Calcul de la statistique du test : $t_{obs} = \dfrac{4,800}{1,117} = 4,296$
6. Décision : $|4,296| \geqslant 2,552$. Nous décidons de rejeter l'hypothèse nulle $(\mathcal{H}_0)$ au seuil de $\alpha = 5\%$.

Contrastes linéaires

- **Objectif :** tester si un groupe de laboratoires est différent d'un autre.
- **Combinaison linéaire des moyennes :** $a_1\mu_1 + a_2\mu_2 + a_3\mu_3 + \cdots + a_I\mu_I$.
- **Contraste linéaire :** combinaison linéaire telle que $a_1 + a_2 + a_3 + \cdots + a_I = 0$
  Exemples :
  1. $\mu_1 - \mu_3$
  2. $1/2(\mu_1 + \mu_2) - \mu_3$
- Un contraste linéaire permet de tester une hypothèse du type :
  « La moyenne des laboratoires 1 et 2 est-elle différente de celle du laboratoire 3 ? »

Test t sur un contraste linéaire

Soit un contraste linéaire $L = a_1\mu_1 + a_2\mu_2 + a_3\mu_3 + \cdots + a_I\mu_I$

1. L'hypothèse nulle :
   $(\mathcal{H}_0) : L = a_1\mu_1 + a_2\mu_2 + a_3\mu_3 + \cdots + a_I\mu_I = 0$
   contre l'hypothèse alternative
   $(\mathcal{H}_1) : L \neq 0$
2. Statistique de test : $L_{obs} = a_1\overline{y}_1 + a_2\overline{y}_2 + a_3\overline{y}_3 + \cdots + a_I\overline{y}_I$
   $t_{obs} = \dfrac{L_{obs}}{s_{L_{obs}}} = \dfrac{L_{obs}}{\sqrt{s_R^2\left(\sum_{i=1}^{I}\dfrac{a_i^2}{n_i}\right)}} \sim t_{n-I}$ sous $(\mathcal{H}_0)$
3. Règle de décision : Accepter $(\mathcal{H}_0)$ si $|t_{obs}| < t_{n-I,1-(\alpha/2)}$

Statistique de rang studentisée

Adaptation du test t pour comparer 2 moyennes a posteriori (n_i supposés égaux)

1. Ordonner les laboratoires en fonction des moyennes observées :
$$\overline{y}_{(1)} \leqslant \overline{y}_{(2)} \leqslant \overline{y}_{(3)} \leqslant \cdots \leqslant \overline{y}_{(I)}$$
2. Puis appliquer la procédure du test qui va suivre

Test basé sur la statistique de rang studentisée

Objectif : Comparer le laboratoire i au laboratoire j, où $i < j$

1. L'hypothèse nulle : $(\mathcal{H}_0) : \mu_i = \mu_j$
contre l'hypothèse alternative : $(\mathcal{H}_1) : \mu_i < \mu_j$
2. Statistique du test : $q_{r,obs} = \dfrac{\overline{y}_{(j)} - \overline{y}_{(i)}}{\sqrt{\dfrac{s_R^2}{J}}}$ avec $r = j - i + 1$
3. Règle de décision : Accepter $(\mathcal{H}_0)$ si $q_{r,obs} < q_{r,n-I}$.
Le seuil critique dépend du nombre de traitements entre i et j et du type d'approche. Dans le cours ici, l'approche sera soit celle de Newman Keuls ou soit celle de Tukey.

Notion de « plus petite différence significative »

- Si nous désirons comparer par une statistique de rang studentisée deux moyennes μ_i et μ_j, nous calculerons la quantité suivante : $q_{r,obs} = \dfrac{\overline{y}_{(j)} - \overline{y}_{(i)}}{\sqrt{\dfrac{s_R^2}{n}}}$ avec $r = j - i + 1$
- Quelle est la plus petite valeur de $\overline{y}_{(j)} - \overline{y}_{(i)}$ à partir de laquelle le test sera rejeté ?
- Réponse : La plus petite valeur de la différence entre les moyennes, à partir de laquelle le test sera rejeté, est égale à : $\overline{y}_{(j)} - \overline{y}_{(i)} \geqslant \sqrt{\dfrac{s_R^2}{n}} \times q_{r,n-I} = W_r$

Test de Newman Keuls

Objectif : Classer les traitements par groupes qui sont significativement différents. La méthode est la suivante :

- **Étape 1 :** Ordonner les moyennes et calculer toutes les différences deux à deux entre moyennes.
- **Étape 2 :** Calculer pour $r = 2$ à I les différences minimum significatives W_r
- **Étape 3 :** Dans le tableau des différences, rechercher toutes les différences significatives en fonction de leur « distance » r
- **Étape 4 :** Classer les traitements par groupes significativement différents.

Test de Tukey

- **But :** Comme pour le test de Newman Keuls, classer les traitements par groupes qui sont significativement différents.
- **Méthode :** Elle est identique à celle du test de Newman-Keuls mais nous prendrons comme différence minimum significative W_I pour toutes les différences. W_I est ici alors noté « HSD » (Honestly Significant Difference)
- **Comparaison des deux méthodes :** La méthode de Tukey trouvera moins de différences significatives que la méthode de Newman Keuls (erreur de type I globale plus faible mais moins de puissance que Newman Keuls)

Contexte du test de Tukey

Les moyennes observées $\overline{y}_i$ sont rangées par ordre croissant. Nous rappelons que nous les notons

$$\overline{y}_{(1)},\ \overline{y}_{(2)},\ldots,\ \overline{y}_{(I)},$$

et les moyennes théoriques associées

$$\mu_{(1)},\ \mu_{(2)},\ldots,\ \mu_{(I)}.$$

Test

La procédure du test de Tukey est la suivante :

Pour chaque $i < i'$, nous considérons **l'hypothèse nulle**

$$(\mathcal{H}_0)\ :\ \mu_{(i)} = \mu_{(i')}$$

contre **l'hypothèse alternative**

$$(\mathcal{H}_1)\ :\ \mu_{(i')} > \mu_{(i)}.$$

Statistique

Nous considérons le rapport :

$$t_{i',i,obs} = \frac{\overline{y}_{(i')} - \overline{y}_{(i)}}{\sqrt{\dfrac{s_R^2}{2}\left(\dfrac{1}{n_{i'}} + \dfrac{1}{n_i}\right)}}. \tag{4}$$

Propriété

Le rapport $t_{i',i,obs}$ défini par (4) est la réalisation d'une variable aléatoire T qui, si l'hypothèse nulle $(\mathcal{H}_0)$ est vraie, suit une loi appelée **étendue studentisée (studentized range)** et que nous notons $\widetilde{\mathcal{T}}_{n-I,I}$.

Décision

Pour un seuil donné α ($= 5\%$ en général), les tables de l'étendue studentisée nous fournissent une valeur critique c telle que $\mathbb{P}_{(\mathcal{H}_0)}[T \leqslant c] = 1 - \alpha$. Alors nous décidons :

$$\begin{cases} \textit{si } c \leqslant t_{i',i,obs} & (\mathcal{H}_1) \textit{ est vraie}, \\ \textit{si } t_{i',i,obs} < c & (\mathcal{H}_0) \textit{ est vraie}. \end{cases}$$

Remarque

La valeur critique c ne dépend que des indices $n - I$, degrés de liberté de la somme des carrés résiduelle, et de I, nombre des moyennes comparées. De plus, les moyennes théoriques, dont les moyennes observées sont comprises entre deux moyennes observées, dont les moyennes théoriques correspondantes sont déclarées égales, sont déclarées égales avec ces dernières.

Sommaire

Le contexte

Des forestiers ont réalisé des plantations d'arbres en trois endroits. Plusieurs années plus tard, ils souhaitent savoir si la hauteur des arbres est identique dans les trois forêts. Chacune des forêts constitue une population. Dans chacune des forêts, nous tirons au sort un échantillon d'arbres, et nous mesurons la hauteur de chaque arbre.

Les données

Forêt 1	Forêt 2	Forêt 3
23, 4	18, 9	22, 5
24, 4	21, 1	22, 9
24, 6	21, 1	23, 7
24, 9	22, 1	24, 0
25, 0	22, 5	24, 0
26, 2	23, 5	
	24, 5	

Le script de **R**

```
>foret<-rep(1:3,c(6,7,5))
>foret
>hauteur<-c(23.4,24.4,24.6,24.9,25.0,26.2,
18.9,21.1,21.1,22.1,22.5,23.5,24.5,22.5,22.9,
23.7,24.0,24.0)
>hauteur
>foret<-factor(foret)
>arbre<-data.frame(foret,hauteur)
>rm(foret)
>rm(hauteur)
>arbre
>str(arbre)
```


Suite du script

```
>moy<-tapply(arbre$hauteur,arbre$foret,mean)
>moy
>moy.g<-mean(arbre$hauteur)
>moy.g
>ecart<-tapply(arbre$hauteur,arbre$foret,sd)
>ecart
>ecart.g<-sd(arbre$hauteur)
>ecart.g
```


Suite du script

```
>plot(arbre$foret,arbre$hauteur)
>points(1:3,moy,pch="@")
>abline(h=moy.g)
```

Suite du script

```
>modele1<-aov(hauteur~foret,data=arbre)
>modele1
>residus<-residuals(modele1)
>residus
>shapiro.test(residus)
>bartlett.test(residus ~ foret,data=arbre)
>summary(modele1)
```


Fin du script

```
>options(contrasts=c("contr.sum",
+"contr.poly"))
>modele2<-lm(hauteur ~ foret,data=arbre)
>modele2
>summary(modele2)
>TukeyHSD(modele1)
```


Les résultats

```
>moy<-tapply(arbre$hauteur,arbre$foret,mean)
>moy
1 2 3
24.75000 21.95714 23.42000
>moy.g<-mean(arbre$hauteur)
>moy.g
[1] 23.29444
```

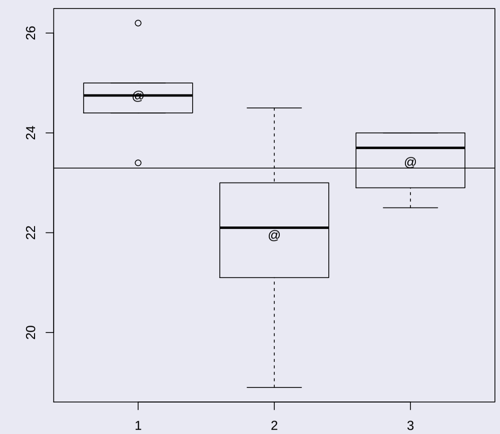

Les résultats

```
>ecart<-tapply(arbre$hauteur,arbre$foret,sd)
>ecart
1 2 3
0.911592 1.824698 0.683374
>ecart.g<-sd(arbre$hauteur)
>ecart.g
[1] 1.737298
```

Les résultats

```
>plot(arbre$foret,arbre$hauteur)
>points(1:3,moy,pch="@")
>abline(h=moy.g)
```


Les résultats



Les résultats

```
>modele1
Call:
aov(formula = hauteur ~ foret, data = arbre)
Terms:
foret Residuals
Sum of Squares 25.30930 26.00014
Deg. of Freedom 2 15
Residual standard error: 1.316565
Estimated effects may be unbalanced
```


Les résultats

```
>shapiro.test(residus)
Shapiro-Wilk normality test
data: residus
W = 0.962, p-value = 0.6399

>bartlett.test(residus~foret,data=arbre)
Bartlett test of homogeneity of variances
data: residus by foret
Bartlett's K-squared = 4.5849, df = 2, p-value
= 0.1010
```

Les résultats

```
>summary(modele1)
Df Sum Sq Mean Sq F value Pr(>F)
foret 2 25.3093 12.6547 7.3007 0.006107
Residuals 15 26.0001 1.7333
```


Les résultats

```
>options(contrasts=c("contr.sum",
+"contr.poly"))
>modele2<-lm(hauteur ~ foret,data=arbre)
>modele2
Call:
lm(formula = hauteur ~ foret, data = arbre)
Coefficients:
(Intercept) foret1 foret2
23.376 1.374 -1.419
```


Les résultats

```
> summary(modele2)
Call:
lm(formula = hauteur ~ foret, data = arbre)
Residuals:
Min 1Q Median 3Q Max
-3.0571 -0.7729 0.1464 0.5707 2.5429
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.3757 0.3133 74.621 < 2e-16
foret1 1.3743 0.4409 3.117 0.00707
foret2 -1.4186 0.4251 -3.337 0.00450
```


Suite des résultats de sorties sous **R**

```
Residual standard error: 1.317 on 15 degrees
of freedom
Multiple R-Squared: 0.4933, Adjusted
R-squared: 0.4257
F-statistic: 7.301 on 2 and 15 DF, p-value:
0.006107
```

Les résultats

```
> TukeyHSD(modele1)
Tukey multiple comparisons of means
95% family-wise confidence level
Fit: aov(formula = hauteur ~ foret, data =
arbre)
$foret
diff lwr upr p adj
2-1 -2.792857 -4.6954236 -0.8902906 0.0045463
3-1 -1.330000 -3.4007541 0.7407541 0.2492545
3-2 1.462857 -0.5395363 3.4652506 0.1735956
```