Analyse de la variance à un facteur

Frédéric Bertrand et Myriam Maumy-Bertrand

IRMA, UMR 7501, Université de Strasbourg

08 juin 2015

Références

Ce cours s'appuie également 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

Sommaire

Sommaire

Sommaire

Objectif

Dans ce chapitre, nous allons étudier un test statistique (nous renvoyons au cours numéro 1 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.

La reproductibilité d'une expérience scientifique est une des conditions qui permettent d'inclure les observations réalisées durant cette expérience dans le processus d'amélioration perpétuelle des connaissances scientifiques. Cette condition part du principe qu'on ne peut tirer de conclusions que d'un événement bien décrit, qui est apparu plusieurs fois, provoqué par des personnes différentes. Cette condition permet de s'affranchir d'effets aléatoires venant fausser les résultats ainsi que des erreurs de jugement ou des manipulations de la part des scientifiques.

Attention

Ne pas confondre cette notion avec la notion de répétabilité.

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

Remarques

1. 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.
2. Il va de soi que lorsque vous rentrerez des données sous un logiciel, vous n'indiquerez pas le mot « Laboratoire » à côté des lettres (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**. Ici le facteur « Laboratoire » est à **effets fixes**.
2. La quantité de Lasalocide. La variable « Lasalocide » est considérée comme quantitative comme généralement tous les résultats obtenus par une mesure. Nous l'appelons **la réponse**.

Notations

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

Pour les observations nous utilisons deux indices :

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

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** de chaque échantillon par :

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

et les **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

Nous allons d'abord importer les données sous **R**, en utilisant les lignes de commande suivantes :

```
> laboratoire<-rep(1:3,c(10,10,10))
> quantite<-c(87,88,84,84,87,81,86,84,88,86,
88,93,88,89,85,87,86,89,88,93,85,84,79,86,81,
86,88,83,83,83)
> jeutotal<-data.frame(laboratoire,quantite)
> moy<-tapply(jeutotal$quantite,
jeutotal$laboratoire,mean)
> moy
> sd<-tapply(jeutotal$quantite,
jeutotal$laboratoire,sd)
> sd
```

Suite de l'exemple

Nous obtenons donc :

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

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 trois conditions suivantes :

1. Elles sont **indépendantes**.
2. Elles sont de **loi gaussienne**.
3. Elles ont **même variance** σ^2 inconnue. C'est la condition d'**homogénéité** ou d'**homoscédasticité**.

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**.

Remarque

En utilisant l'autre modèle, ce test permet de tester l'hypothèse nulle

$$\mathcal{H}_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$$

contre l'hypothèse alternative

$$\mathcal{H}_1 : \text{il y a au moins un des } \alpha_i \text{ qui est non nul.}$$

Sommaire

Remarque

Le test de Fisher est fondé sur deux propriétés : une propriété qui porte sur les moyennes et une autre sur les variances.

Première propriété

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

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

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 :

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

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 :

$$SC_{tot} = SC_F + SC_{res}. \tag{2}$$

Retour à l'exemple

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

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

et d'autre part :

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

Donc lorsque nous faisons la somme des deux derniers résultats 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 :

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

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_{res}.

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_{res}.

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

$$CM_{res} = \frac{SC_{res}}{n-I}.$$

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^2_{res} lorsqu'il sert à estimer la variance des erreurs.

Sa valeur observée sur l'échantillon est ainsi notée cm_{res} ou s^2_{res}.

Propriété

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

$$F_{obs} = \frac{cm_F}{cm_{res}}$$

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 et conclusion du test

Pour un seuil donné α (=5%=0,05 en général), les tables de Fisher nous fournissent une valeur critique c_α telle que $\mathbb{P}_{\mathcal{H}_0}(F \leqslant c_\alpha) = 1 - \alpha$. Si la valeur de la statistique calculée sur l'échantillon, notée F_{obs}, est supérieure ou égale à c_α, alors le test est significatif. Vous rejetez $\mathcal{H}_0$ et vous décidez que $\mathcal{H}_1$ est vraie avec un risque d'erreur de première espèce $alpha = 5\%$. Si la valeur de la statistique calculée sur l'échantillon, notée F_{obs}, est strictement inférieure à c_α, alors le test n'est pas significatif. Vous conservez $\mathcal{H}_0$ avec un risque d'erreur de deuxième espèce β qu'il faut évaluer.

Tableau de l'ANOVA

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

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

Retour à l'exemple

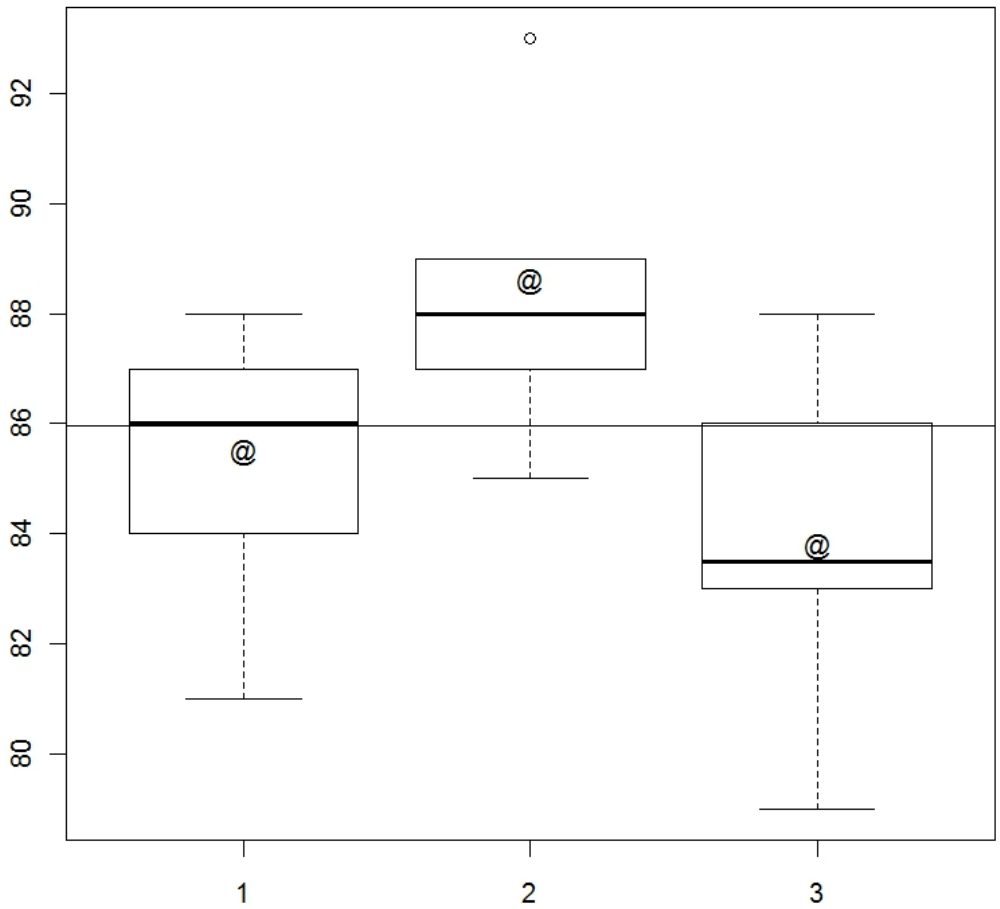
Pour les données de l'exemple des laboratoires, le tableau de l'analyse de la variance s'écrit :

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			

Décision et conclusion du test

Pour un seuil $\alpha = 5\%$, les tables de Fisher nous fournissent la valeur critique $F_c = 3,35$. Le test est significatif puisque $9,49 \geqslant 3,35$. Nous décidons donc de rejeter l'hypothèse nulle $\mathcal{H}_0$ 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.


92
90
88
86
84
82
80
@
@
@
1
2
3

Remarques

1. Nous avons décidé que les moyennes théoriques sont différentes dans leur ensemble, mais nous aurions très bien pu trouver le contraire.
2. 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** 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 :

$$\begin{cases} \text{si } r_{obs} \leqslant c & \text{alors le test est significatif. } \mathcal{H}_1 \text{ est acceptée.} \\ \text{si } c < r_{obs} & \text{alors le test n'est pas significatif. } \mathcal{H}_0 \text{ est acceptée.} \end{cases}$$

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 $c < r_{obs}$, le test n'est pas significatif. Nous décidons de ne pas rejeter l'hypothèse nulle $\mathcal{H}_0$. Le risque d'erreur associé à cette décision est un risque de deuxième espèce β que nous ne pouvons pas évaluer. **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 (0,6431) est supérieure à 0,05, le test n'est pas significatif. Nous décidons de ne pas rejeter l'hypothèse nulle $\mathcal{H}_0$. Le risque d'erreur associé à cette décision est un risque de deuxième espèce β que nous ne pouvons pas évaluer. **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

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

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.

Remarque

Ce test dépend de la normalité des résidus. Il se fait donc après avoir vérifié la normalité des résidus.

Décision et conclusion du test

Pour un seuil donné α ($= 5\%$ en général), les tables du khi-deux nous fournissent une valeur critique c_α telle que $\mathbb{P}_{\mathcal{H}_0}(B \leqslant c_\alpha) = 1 - \alpha$. Si la valeur de la statistique calculée sur l'échantillon, notée B_{obs}, est supérieure ou égale à c_α, alors le test est significatif. Vous rejetez $\mathcal{H}_0$ et vous décidez que $\mathcal{H}_1$ est vraie avec un risque d'erreur de première espèce $\alpha = 5\%$. Si la valeur de la statistique calculée sur l'échantillon, notée B_{obs}, est strictement inférieure à c_α, alors le test n'est pas significatif. Vous conservez $\mathcal{H}_0$ avec un risque d'erreur de deuxième espèce β que vous ne pouvez pas évaluer.

Retour à l'exemple

Nous obtenons avec le logiciel **R** et la commande `bartlett.test` :

```
> bartlett.test(residuals(modele)~laboratoire,
+data=analyse)
Bartlett test of homogeneity of variances
data: residuals(modele) by laboratoire
Bartlett's K-squared = 0.3024, df = 2, p-value = 0.8597
```

Retour à l'exemple : suite

En se souvenant que **les n_i sont tous égaux**, nous lisons, 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 $B_{obs} < c$, le test n'est pas significatif. Nous décidons de ne pas rejeter l'hypothèse nulle $\mathcal{H}_0$. Le risque d'erreur associé à cette décision est un risque de deuxième espèce β que nous ne pouvons pas évaluer. **Nous décidons 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.
  Méthodes :
  - Méthode de Bonferroni,
  - Méthode des contrastes linéaires.

Exemple

Montrer que les deux premiers laboratoires sont différents du dernier.

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.
  Méthodes :
  - Méthode basée sur la statistique d'écart studentisée,
  - Méthode de Tukey

Exemple

Prendre la plus grande et la plus petite moyenne dans l'exemple des laboratoires et tester si elles sont vraiment différentes.

Correction de Bonferroni

Idée

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

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

I	$n_{\mathcal{C}}$	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

Test t de Bonferroni ou test de Dunn

Objectif : comparer deux à deux toutes les moyennes possibles des I groupes.

1. Calcul du nombre de comparaisons : $n_{\mathcal{C}} = (I \times (I-1))/2$
2. Erreur de type I globale : $\alpha = 5\% = 0,05$
3. Erreur pour chaque test : $\alpha' = 0,05/n_{\mathcal{C}}$
4. Hypothèses : $\mathcal{H}_0 : \mu_i = \mu_j$ contre $\mathcal{H}_1 : \mu_i \neq \mu_j$
5. Calcul de la statistique du 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. Décision : si $|t_{obs}| < t_{n-I;1-(\alpha'/2)}$, alors le test n'est pas significatif. Si $|t_{obs}| \geqslant t_{n-I;1-(\alpha'/2)}$, alors le test est significatif

Retour à l'exemple

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

1. Calcul du nombre de comparaisons : $n_{\mathcal{C}} = (3 \times 2)/2 = 3$
2. Erreur de type I globale : $\alpha = 5\% = 0,05$
3. Erreur pour chaque 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 : comme $|4,296| \geqslant 2,552$, le test est significatif. Nous décidons de rejeter l'hypothèse nulle $\mathcal{H}_0$ et d'accepter l'hypothèse alternative $\mathcal{H}_1$ au seuil de $\alpha = 5\%$.

Méthode des 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 $a_1 + a_2 + a_3 + \cdots + a_I = 0$

Exemple

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. Hypothèses : $\mathcal{H}_0 : L = 0$ contre $\mathcal{H}_1 : L \neq 0$
2. Calcul de la statistique du test :
   $L_{obs} = a_1\overline{y}_1 + a_2\overline{y}_2 + a_3\overline{y}_3 + \cdots + a_I\overline{y}_I$
   $$t_{obs} = \frac{L_{obs}}{s_{L_{obs}}} = \frac{L_{obs}}{\sqrt{s_R^2\left(\sum_{i=1}^{I}\frac{a_i^2}{n_i}\right)}} \sim t_{n-I} \text{ sous } \mathcal{H}_0$$
3. Règle de décision : si $|t_{obs}| < t_{n-I,1-(\alpha/2)}$, alors le test n'est pas significatif. Si $|t_{obs}| \geqslant t_{n-I,1-(\alpha/2)}$, alors le test est significatif.

Statistique d'écart studentisée

Adaptation du test t pour comparer deux moyennes a posteriori.

- Ordonner les I groupes en fonction des moyennes observées :

$$\overline{y}_{(1)} \leqslant \overline{y}_{(2)} \leqslant \overline{y}_{(3)} \leqslant \cdots \leqslant \overline{y}_{(I)}.$$

- Puis appliquer la procédure du test qui va suivre.

Remarque

Attention : dans la suite, les formules seront données pour le cas où les tailles n_i sont supposées égales.

Test basé sur la statistique d'écart studentisée

Objectif : comparer le groupe i au groupe j, où $i < j$

1. Hypothèses : $\mathcal{H}_0 : \mu_i = \mu_j$ contre $\mathcal{H}_1 : \mu_i < \mu_j$
2. Calcul de la 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 : si $q_{r,obs} < q_{r,n-I,1-(\alpha/2)}$, alors le test n'est pas significatif. Si $q_{r,obs} \geqslant q_{r,n-I,1-(\alpha/2)}$, alors le test est significatif.

Question

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 : $\overline{y}_{(j)} - \overline{y}_{(i)} \geqslant \sqrt{\frac{s_R^2}{n}} \times q_{r,n-I,1-(\alpha/2)}$.

Contexte du test de Tukey

Les moyennes observées $\overline{y}_i$ sont rangées par ordre croissant. Nous rappelons que nous les notons par : $\overline{y}_{(1)}, \ \overline{y}_{(2)}, \ldots, \ \overline{y}_{(I)}$, et les moyennes théoriques associées par : $\mu_{(1)}, \ \mu_{(2)}, \ldots, \ \mu_{(I)}$.

Remarques

1. Le test de Tukey est également appelé test de la différence franchement significative (HSD : Honestly Significant Difference).
2. Le test s'appuie sur la statistique d'écart studentisée q pour les comparaisons, si ce n'est que q_{HSD} correspond toujours à la valeur maximale de q_r.

Test de Tukey

1. Hypothèses : $\mathcal{H}_0 : \mu_{(i)} = \mu_{(i')}$ contre $\mathcal{H}_1 : \mu_{(i')} > \mu_{(i)}$.
2. Calcul de la statistique du test :

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

   La variable $t_{i',i,obs}$ est la réalisation d'une v.a. T qui, si $\mathcal{H}_0$ est vraie, suit une loi appelée **étendue studentisée (studentized range)** et que nous notons $\widetilde{\mathcal{T}}_{n-I,I}$.
3. Règle de décision : si $t_{i',i,obs} \geqslant c$, alors le test est significatif. Vous rejetez $\mathcal{H}_0$ et vous décidez que $\mathcal{H}_1$ est vraie avec un risque d'erreur de première espèce $\alpha = 5\%$. Si $\leqslant t_{i',i,obs} < c$, alors le test n'est pas significatif. Vous conservez $\mathcal{H}_0$ avec un risque d'erreur de deuxième espèce β qu'il faut évaluer.

Retour à l'exemple

```
> modele1<-aov(lasalocide~laboratoires,
+data=analyse)
> TukeyHSD(modele1)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = lasalocide ~ laboratoires, data = analyse)

$laboratoires
    diff        lwr       upr     p adj
2-1  3.1  0.3299809  5.870019 0.0259501
3-1 -1.7 -4.4700191  1.070019 0.2969093
3-2 -4.8 -7.5700191 -2.029981 0.0005724
```

Groupement avec la méthode de Tukey

Laboratoire	Taille	Moyenne	Groupement	
A	10	$85,5$	B	
B	10	$88,6$		A
C	10	$83,8$	B	

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.
De plus, des études ont montré que la hauteur des arbres suit une loi normale.

Les données

Forêt 1	Forêt 2	Forêt 3
25, 2	22, 6	23, 4
24, 7	22, 1	23, 9
24, 6	23, 3	23, 7
24, 8	21, 7	24, 2
24, 0	23, 5	24, 0
25, 8	22, 5	23, 1
25, 5	21, 6	24, 5
26, 1	22, 7	24, 3
24, 5	21, 3	
25, 3	21, 5	
	22, 2	
	22, 4	

Le script de **R**

```
>foret<-rep(1:3,c(10,12,8))
>foret
>hauteur<-c(25.2,24.7,24.6,24.8,24.0,25.8,25.5,
26.1,24.5,25.3,22.6,22.1,23.3,21.7,23.5,22.5,
21.6,22.7,21.3,21.5,22.2,22.4,23.4,23.9,23.7,
24.2,24.0,23.1,24.5,24.3)
>hauteur
>foret<-factor(foret)
>arbre<-data.frame(foret,hauteur)
>rm(foret)
>rm(hauteur)
>arbre
>str(arbre)
```

Suite du script de **R**

```
>moyenne<-tapply(arbre$hauteur,arbre$foret,mean)
>moyenne
>moyenne.gene<-mean(arbre$hauteur)
>moyenne.gene
>ecart<-tapply(arbre$hauteur,arbre$foret,sd)
>ecart
>ecart.g<-sd(arbre$hauteur)
>ecart.g
>boxplot(arbre$hauteur~arbre$foret)
>points(1:3,moy,pch="@")
>abline(h=moy.g)
```

Suite du script de **R**

```
>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 de **R**

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

Les résultats de sorties sous **R**

```
>moyenne<-tapply(arbre$hauteur,arbre$foret,mean)
>moyenne
       1        2        3
25.05000 22.28333 23.88750
>moyenne.gene<-mean(arbre$hauteur)
>moyenne.gene
[1] 23.63333
```

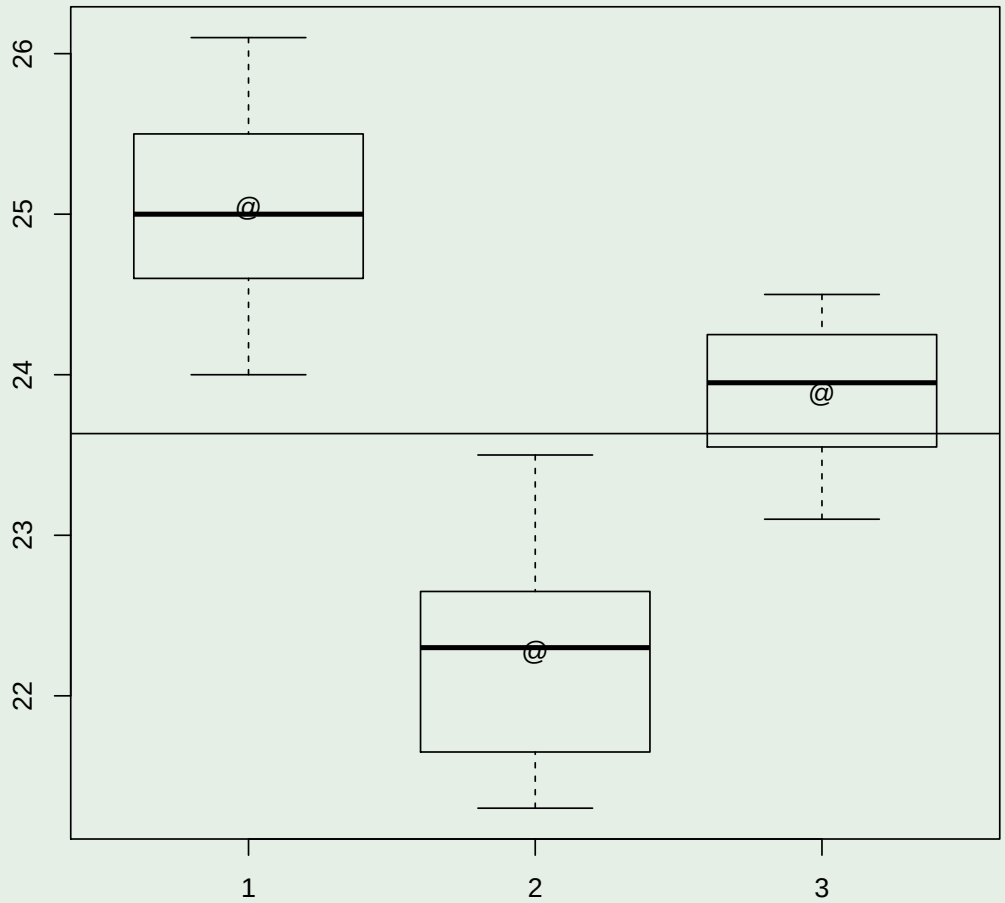
Les résultats de sorties sous **R**

```
>ecart<-tapply(arbre$hauteur,arbre$foret,sd)
>ecart
        1         2         3
0.6450667 0.6926016 0.4703722
>ecart.g<-sd(arbre$hauteur)
>ecart.g
[1] 1.352223
```

Les résultats de sorties sous **R**

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

Suite des résultats de sorties sous **R**



Les résultats de sorties sous **R**

```
>modele1
Call:
   aov(formula=hauteur~foret,data=arbre)

Terms:
                     foret  Residuals
Sum of Squares   42.45625   10.57042
Deg. of Freedom          2          27

Residual standard error: 0.6256971
Estimated effects may be unbalanced
```

Les résultats de sorties sous **R**

```
>shapiro.test(residus)

   Shapiro-Wilk normality test
data:  residus
W = 0.9779, p-value = 0.7671

>bartlett.test(residus~foret,data=arbre)

   Bartlett test of homogeneity of variances
data:  residus by foret
Bartlett's K-squared = 1.152, df = 2,
p-value = 0.5622
```

Les résultats de sorties sous **R**

```
>summary(modele1)
             Df  Sum Sq Mean Sq F value   Pr(>F)
foret         2   42.46  21.228   54.22  3.5e-10
Residuals    27   10.57   0.391
```

Les résultats de sorties sous **R**

```
>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
    23.740        1.310       -1.457
```

Les résultats de sorties sous **R**

```
> summary(modele2)
Call:
lm(formula=hauteur~foret,data=arbre)
Residuals:
    Min      1Q  Median      3Q     Max
-1.0500 -0.4781  0.0625  0.3885  1.2167
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  23.7403     0.1158 204.990  < 2e-16
foret1        1.3097     0.1627   8.051 1.19e-08
foret2       -1.4569     0.1558  -9.349 5.90e-10
```

Les résultats de sorties sous **R**

```
Residual standard error: 0.6257 on 27
degrees of freedom
Multiple R-squared: 0.8007,
Adjusted R-squared: 0.7859
F-statistic: 54.22 on 2 and 27 DF,
p-value: 3.504e-10
```

Les résultats de sorties sous **R**

```
> 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 -2.766667 -3.4309213 -2.1024120 0.000000
3-1 -1.162500 -1.8983768 -0.4266232 0.001549
3-2  1.604167  0.8960689  2.3122645 0.000017
```

Groupement avec la méthode de Tukey

Forêt	Taille	Moyenne	Groupement		
1	8	25,05000	*A*		
2	12	22,28333		*B*	
3	10	23,88750			*C*