Chapitre 5

Comparaisons multiples

Une fois que toutes les hypothèses d'une ANOVA ont été vérifiée et que l'analyse a été effectuée, deux conclusions sont possibles, soit on rejette H_0, soit on n'a pas assez de preuves pour le faire.

Dans le dernier cas, généralement l'analyse s'arrête là. On conclut qu'il n'y a pas de différences significatives entre les groupes. Cependant, dans le premier cas, l'hypothèse d'égalité des groupes est écartée. On veut identifier les modalités du facteur qui sont responsables du résultat significatif. On veut parfois classer les moyennes observées et identifier les différences significaives pour toutes les paires de moyennes.

Dans ce chapitre, on s'intéresse au problème de comparaisons multiples des moyennes $\mu_1, \mu_2, \cdots, \mu_I$ ou de combinaisons linéaires de ces moyennes. On considère également la construction d'intervalles de confiances simultanés pour un ensemble de différences, ou de combinaisons linéaires de moyennes.

5.1 Mise en situation: seuil local et seuil global

Supposons que toutes les tailles d'échantillons sont égales (i.e. $n_i = n$, $i = 1, \ldots, n$). On dit alors que l'expérience est balancée. Dans une ANOVA à un facteur on teste d'abord une hypothèse globale d'homogénéité, $H_0^g : \mu_1 = \mu_2 = \cdots = \mu_I$ versus H_1 : il existe $i \neq j$ tels que $\mu_i \neq \mu_j$.

Souvent, on est également intéressé à comparer toutes les moyennes deux à deux. Les hypothèses nulles sont alors $H_0^{ij} : \mu_i = \mu_j$. Evidemment le test global qui compare toutes les moyennes d'un coup et l'ensemble des tests "locaux" qui compare les moyennes deux à deux ne sont pas équivalents. Pour comparer des moyennes deux à deux au seuil α on peut rejeter H_0^{ij} si la statistique t suivante satisfait

$$t_{obs}^{ij} = \left| \frac{\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}}{\sqrt{2\hat{\sigma}^2/n}} \right| > t_{I(n-1),\alpha/2},$$

où $\hat{\sigma}^2 = MSE = SSW/\{I(n-1)\}$. En fait les moyennes $\bar{Y}_{i\cdot}$ et $\bar{Y}_{j\cdot}$ sont déclarées différentes si la valeur absolue de leur différence est supérieur au Least Significant Difference (LSD) de Fisher où

$$\text{LSD} = t_{I(n-1),\alpha/2}\sqrt{2\hat{\sigma}^2/n}.$$

Pour comprendre la différence entre le test global et les tests deux à deux, on va simuler des données où H_0^g est vraie et on va faire le test global et l'ensemble des tests deux à deux. On prend $I = 10$ et $n = 5$. Il y a donc $10 \times 9/2 = 45$ paires de moyennes différentes. La simulation utilise les énoncés SAS suivants:

```
/*Simulation de 10 echantillons de taille 5 de v.a. N(0,1)*/
 data simul;
    do i=1  to 10;
    do j=1  to 5;
    Y=rannor(2456);
    output;
end;
end;
run;
 /* Analyse de variance a un facteur*/
proc glm data=simul;
class i;
model y=i;
means i /lsd;
```

```
run;
```

TABLE ANOVA

Source	DF	Sum of Squares	Mean Square	F Value	Pr> F
Model	9	9.66833	1.07426	1.44	0.2055
Error	40	29.9102	0.74775		
Corrected Total	49	39.5785			

On note que le test F de la table ANOVA n'est pas significatif au seuil 5%. Regardons maintenant les comparaisons des 10 moyennes entre elles. La sortie SAS pour cette énoncé rappelle d'abord les éléments sous-jacents au calcul de la LSD. Une fois que ce calcul est fait les moyennes sont mises en ordre décroissant et les résultats des 45 tests sont synthétisés de la manière suivante: si deux moyennes sont jointes par une même ligne verticale, elles ne sont pas significativement différentes. Le nombre de lignes verticales donne le nombre de groupes de moyennes qui ne sont pas significativement différentes.

```
 t Tests (LSD) for Y
NOTE: This test controls the Type I comparisonwise error rate, not
 the experimentwise error rate.
```

Alpha	0.05
Error Degrees of Freedom	40
Error Mean Square	0.747754
Critical Value of t	2.02108
Least Significant Difference	1.1053

```
Means with the same letter are not significantly different.
```

t	Grouping		Mean	N	i
		A	0.7655	5	4
		A			
	B	A	0.3492	5	5
	B	A			
	B	A	0.2365	5	6
	B	A			
	B	A	0.1975	5	3
	B	A			
	B	A	0.0641	5	7
	B	A			
	B	A	-0.0525	5	8
	B	A			
	B	A	-0.1939	5	2
	B				
	B		-0.3744	5	9
	B				
	B		-0.6723	5	10
	B				
	B		-0.7141	5	1

Ce tableau souligne que 3 des 45 tests (3/45=6.7%) de comparaisons deux à deux sont signifiquatifs au seuil 5%, ceux pour les paires (4,10), (4,9) et (4,1).

Pour expliquer l'apparente contradiction entre le test global (les moyennes ne sont pas significativement différentes à 5%) et les tests deux à deux (3 paires de moyennes sont significativement différents à 5%). Il faut rappeler la définition du seuil. C'est la probabilité de rejeter H_0 lorsque cette dernière est vraie. Si les moyennes sont toutes égales lorsque l'on fait 45 tests de comparaison deux à deux on devrait avoir en moyenne $45 \times 0.05 = 2.25$

tests significatifs même si les moyennes sont toutes égales. En fait le seuil que rejette H_0^g lorsqu'au moins une différence de moyennes est supérieure à LSD est une mauvaise procédure statistique car son seuil est proche de 1.

Cet exemple introduit la notion de seuil global (experimentwise) et local (componentwise) et montre qu'il est très important de contrôles le seuil global lorsque l'on rejette H_0^g. La problématique local vs global se retrouve également lorsque l'on construit des intervalles de confiance. En effet on sait que l'intervalle de confiance à 95% pour $\mu_1 - \mu_2$ est $\bar{Y}_{1.} - \bar{Y}_{2.} \pm LSD$. Comment construire B tel que l'ensemble des $I(I-1)/2$ intervalles de confiance $\{Y_{i.} - \bar{Y}_{j.} \pm B : 1 \leq i < j \leq I\}$ ait un niveau de confiance simultané de 95%. Si on prend $B = LSD$, le noveau de confiance simultané sera beaucoup plus petit que 95%. En fait pour constuire B il faut utiliser la distribution du "Studentized range" de Tukey.

5.2 Combinaisons linéaires et contrates

Soient K combinaisons linéaires $L_1, L_2, \cdots, L_K$ définies par

$$L_k = \sum_{i=1}^{I} c_{ki}\mu_i \text{ pour } k = 1, 2, \cdots, K$$

où les c_{ki} sont des constantes réelles. On dit que les L_k sont des contrastes si $\sum_{i=1}^{I} c_{ki} = 0$. Ce type de combinaison est souvent utilisé pour interpréter un résultat significatif.

EXEMPLE **5.1** *Un agriculteur est en train de comparer 4 types d'engrais: E_1, E_2, E_3 et E_4. Il effectue un test d'ANOVA global, en prenant soin bien sur de vérifier les hypothèses du modèle. Il rejette H_0. Il s'intéresse maintenant à la différence entre le premier et le dernier engrais d'une part et entre le premier et la moyenne des trois derniers d'autre part. Ici, on a $K = 2$, $L_1 = \mu_1 - \mu_4$ et $L_2 = \mu_1 - (\mu_2 + \mu_3 + \mu_4)/3$. Les constantes c_{ki} sont alors égales à $\{c_{11}, c_{12}, c_{13}, c_{14}, c_{21}, c_{22}, c_{23}, c_{24}\} = \{1, 0, 0, -1, 1, -1/3, -1/3, -1/3\}$.*

On a vu à la section (3.8) que pour chaque $k = 1, 2, \cdots, K$, l'intervalle de confiance au niveau $1 - \alpha$ pour L_k s'écrit

$$I_k = \left[\sum_{i=1}^{I} c_{ki}\bar{Y}_{i.} - t_{N-I,\alpha/2}\sqrt{MSE\sum_{i=1}^{I}\frac{c_{ki}^2}{n_i}}, \sum_{i=1}^{I} c_{ki}\bar{Y}_{i.} + t_{N-I,\alpha/2}\sqrt{MSE\sum_{i=1}^{I}\frac{c_{ki}^2}{n_i}}\right]$$

On a alors

$$P[L_k \in I_k] = 1 - \alpha \text{ pour chaque } k = 1, 2, \cdots, K.$$

Or en pratique, plusieurs applications s'intéressent à obtenir des intervalles de confiances simultanés, c'est à dire vérifiant

$$P[L_k \in I_k \text{ pour chaque } k = 1, 2, \cdots, K] = 1 - \alpha.$$

Ces probabilités ne sont pas équivalentes. Dans ce chapitre, on présente plusieurs méthodes pour construire de tels intervalles de confiance.

5.3 Méthode de Tukey

Cette méthode requiert que les tailles d'échantillons soient égales. Soit n la taille commune. Elle permet de trouver des intervalles de confiance simultanées pour toutes les différences $\mu_i - \mu_j$, au nombre de $\binom{I}{2} = I(I-1)/2$. Avant d'exposer cette méthode, on a besoin d'un petit développement théorique.

Soient $Z_1, Z_2, \cdots, Z_I$ des variables aléatoires i.i.d. suivant la loi normale standard $\mathrm{N}(0, 1)$. Posons

$$R = Z_{(I)} - Z_{(1)}$$

où $Z_{(1)}$ et $Z_{(I)}$ sont deux statistiques d'ordre égales au minimum et maximum, respectivement, de $\{Z_1, Z_2, \cdots, Z_I\}$.

Définition **5.1** *La variabe R ainsi définie est appelée l'étendue de l'échantillon $\{Z_1, Z_2, \cdots, Z_I\}$. Sa distribution, appelée distribution de l'étendue gaussienne, ne dépend que de I et est parfois notée Q_I*

Soient maintenant deux variables aléatoires indépendantes U et V telles que $U \sim Q_I$ et $V \sim \chi^2_k$. Posons

$$S = \frac{U}{\sqrt{V/k}}$$

DÉFINITION **5.2** *La distribution de S est appelée la distribution de l'étendue gaussienne studentisée. Cette distribution, qui ne dépend que de I et k, est parfois notée $Q_{I,k}$.*

Revenons maintenant à l'analyse de la variance avec des tailles échantillons égales. Pour $i = 1, 2, \cdots, I$, posons $Z_i = (\bar{Y}_{i.} - \mu_i)/(\sqrt{\sigma^2/n})$. Les Z_i sont des variables aléatoires indépendantes distribuées selon la loi normale standard. On a donc

$$U = Z_{(I)} - Z_{(1)} \sim Q_I$$

D'autre part, posons $V = (N - I)MSE/\sigma^2$. On sait que $V \sim \chi^2_{N-I}$ et U et V sont indépendantes. On a alors

$$\frac{U}{\sqrt{V/(N-I)}} \sim Q_{I,N-I}$$

Ce résultat sécrit alors

$$\frac{\max_{1\leq i\leq I}(\bar{Y}_{i.} - \mu_i) - \min_{1\leq i\leq I}(\bar{Y}_{i.} - \mu_i)}{\sqrt{MSE/n}} \sim Q_{I,N-I}.$$

Ce résultat nous aide alors à construire des intervalles de confiance simultanés pour $\mu_i - \mu_j$. En effet, on a:

$$\begin{aligned}
1-\alpha &= P\{\frac{\max_{1\leq i\leq I}(\bar{Y}_{i.} - \mu_i) - \min_{1\leq i\leq I}(\bar{Y}_{i.} - \mu_i)}{\sqrt{MSE/n}} \leq Q_{I,N-I,\alpha}\} \\
&= P\{\max_{1\leq i,j\leq I}|\frac{(\bar{Y}_{i.} - \mu_i) - (\bar{Y}_{j.} - \mu_j)}{\sqrt{MSE/n}}| \leq Q_{I,N-I,\alpha}\} \\
&= P\{\cap_{i\leq i,j\leq I}(|\frac{(\bar{Y}_{i.} - \mu_i) - (\bar{Y}_{j.} - \mu_j)}{\sqrt{MSE/n}}| \leq Q_{I,N-I,\alpha})\} \\
&= P\{\cap_{i\leq i,j\leq I}(|\frac{(\bar{Y}_{i.} - \bar{Y}_{j.}) - (\mu_i - \mu_j)}{\sqrt{MSE/n}}| \leq Q_{I,N-I,\alpha})\}
\end{aligned}$$

On en déduit que les intervalles de confiance

$$I_{i,j} = [(\bar{Y}_{i.} - \bar{Y}_{j.}) - Q_{I,N-I,\alpha}\sqrt{\frac{MSE}{n}}, (\bar{Y}_{i.} - \bar{Y}_{j.}) + Q_{I,N-I,\alpha}\sqrt{\frac{MSE}{n}}]$$

sont de niveau simultané $1 - \alpha$ pour les différences $\mu_i - \mu_j$

REMARQUE **5.1** *La quantité* $Q_{I,N-I,\alpha}\sqrt{\frac{MSE}{n}}$ *est appelée plus petite différence significative ou (PPDS). En anglais on l'appelle TSD pour Tukey Significant Difference. Elle est égale à la demi-longueur de l'intervalle de confiance défini ci-haut. Ces* $I(I-1)/2$ *intervalles de confiance simultanés de Tukey sont équivalents aux* $I(I-2)/2$ *tests bilatéraux* $H_{0,ij}$: $\mu_i - \mu_j = 0$ *au seuil global* α. *On rejette* H_0^{ij} *si* 0 *n'appartient pas à* $I_{i,j}$, *c'est à dire si* $|\bar{Y}_{i.} - \bar{Y}_{j.}| > PPDS$.

Note: Pour réanalyser les données simulées à la section 5.1 en utilisant la méthode de Tukey pour faire les comparaisons multiples, il suffit de mettre `means i /tukey;` comme option dans GLM. On obtient comme plus petite différence significative TSD=1.8309. La plus grande différence entre deux moyennes est de 1.48. La méthode de Tukey ne détecte donc aucune différence significative. Elle donne le même résultat que le test global F. Evidemment les deux tests sont différents; cependant puisque les deux utilisent le même seuil global il donnent souvent le même résultat.

5.4 Méthode de Bonferroni

Soient K événements $A_1, A_2, \cdots, A_K$. On sait que le complément de l'union $\cup_{1\leq k\leq K} A_k$ est l'intersection des compléments $\cap_{1\leq k\leq K} A_k^c$, d'où

$$\begin{aligned} P(\cap_{1\leq k\leq K} A_k) &= 1 - P((\cap_{1\leq k\leq K} A_k)^c) \\ &= 1 - P(\cup_{1\leq k\leq K} A_k^c) \\ &\geq 1 - \sum_{k=1}^{K} P(A_k^c) \end{aligned}$$

Pour obtenir des intervalles de confiance simultanés $I_1, I_2, \cdots, I_K$ pour $L_1, L_2, \cdots, L_K$ au niveau global $1-\alpha$, la méthode de Bonferroni consiste à construire K intervalles de confiance au niveau $1-\alpha/K$, c'est à dire vérifiant $P[L_k \in I_k] = 1 - \alpha/K$. Le niveau de confiance simultané des intervalles $I_1, I_2, \cdots, I_K$ est égal au moins à α. En effet, on a

$$P[L_k \in I_k \text{ pour chaque } k = 1, 2, \cdots, K] = P(\cap_{k=1}^{K}(L_k \in I_k))$$

$$\begin{aligned} &\geq 1 - \sum_{k=1}^{K} P[L_k \notin I_k] \\ &= 1 - \sum_{k=1}^{K} \frac{\alpha}{K} = 1 - \alpha \end{aligned}$$

Cette méthode a l'avantage de pouvoir construire des intervalles de confiance pour n'importe quelle combinaison linéaire des $\mu_1, \mu_2, \cdot, \mu_I$ et pas seulement les différences $\mu_i - \mu_j$. Aussi, elle ne requière pas que les tailles déchantillons soient égales et peut s'appliquer à d'autres modèles que l'ANOVA. Cependant, elle ne fonctionne bien que lorsque le nombre de comparaisons K à faire n'est pas très grand. En effet, si K devient grand, le rapport α/K devient petit et la largueur des intervalles devient grande car elle est proportionelle à $t_{N-I,\alpha/2K}$. Généralement, lorsque les tailles des échantillons sont égales et qu'on s'intéresse aux différences de moyennes, la méthode de Tukey est plus précise.

REMARQUE **5.2** *Dans le cas d'une différence $\mu_i - \mu_j$, la largueur de l'intervalle de confiance est $2 \times t_{N-I,\alpha/2K}\sqrt{MSE(\frac{1}{n_i} + \frac{1}{n_j})}$. Dans ce cas, la PPDS dépend du couple (i, j) à travers les tailles d'échantillons et est égale à $PPDS = t_{N-I,\alpha/2K}\sqrt{MSE(\frac{1}{n_i} + \frac{1}{n_j})}$.*

Dans l'exemple simulée la PPDS construite avec la méthode de Bonferroni pour préserver le seuil global des 45 tests de comparaison de paires de moyennes est de $t_{40,1-0.05/90}\sqrt{2\hat{\sigma}^2/5} = 1.92$. C'est évidemment plus grand que le TSD=1.83. Dans ce cas la méthode de Tukey est toujours meilleur que celle de Bonferronni car elle s'assure que le seuil global est égal au seuil désiré. La méthode de Bonferronni nous assure que le seuil global est inférieur au seuil souhaité. Il s'agit d'un alternative intéressante si on s'intéresse seulement à un sous-ensemble des $I(I-1)/2$ paires de moyennes.

5.5 La méthode de Scheffé

La méthode de Scheffé est utilisée pour donner des intervalles de confiance simultanés pour TOUTES les combinaisons linéaires. C'est donc une méthode très générale et à cause de cette généralité, elle est moins puissante pour détecter des différences significatives.

PROPOSITION **5.1** *Avec les définitions usuelles d'ANOVA, on a:*

$$P[L_k \in I_k \text{ pour chaque toutes les combinaisons linéaires}] = 1 - \alpha$$

où

$$I_k = \left[\sum_{i=1}^{I} c_{ki} \bar{Y}_{i.} - \sqrt{IF_{I,N-I,\alpha} MSE \sum_{i=1}^{I} \frac{c_{ki}^2}{n_i}}, \sum_{i=1}^{I} c_{ki} \bar{Y}_{i.} + \sqrt{IF_{I,N-I,\alpha} MSE \sum_{i=1}^{I} \frac{c_{ki}^2}{n_i}} \right]$$

Dans certaines situations, on se limite aux contrastes, c'est à dire aux combinasons linéaires L_k vérifiant $\sum_{i=1}^{I} c_{ki} = 0$. Dans ce cas les intervalles de confiance simultanés deviennent:

$$I_k = \left[\sum_{i=1}^{I} c_{ki} \bar{Y}_{i.} - \sqrt{(I-1)F_{I-1,N-I,\alpha} MSE \sum_{i=1}^{I} \frac{c_{ki}^2}{n_i}}, \sum_{i=1}^{I} c_{ki} \bar{Y}_{i.} + \sqrt{(I-1)F_{I-1,N-I,\alpha} MSE \sum_{i=1}^{I} \frac{c_{ki}^2}{n_i}} \right]$$

REMARQUE **5.3** *La plus petite différence significative $PPDS$ pour la différences $\mu_i - \mu_j$ est égale selon cette méthode à:*

$$PPDS = \sqrt{(I-1)F_{I-1,N-I,\alpha} MSE (\frac{1}{n_i} + \frac{1}{n_j})}$$

Dans l'exemple simulé, la méthode de Scheff donne une $PPDS = \sqrt{9 \times F_{9,40,0.05} \times 2\hat{\sigma}^2/5} = 2.39$. C'est la plus grande valeur obtenue. Ceci explique que l'on utilise à peu près jamais cette méthode pour classer des moyennes.

5.6 Synthèse et conclusion

Notons d'abord que toutes les méthodes vues ici peuvent accomoder une expérience non balancée, même la méthode de Tukey qui devient celle de Tukey Kramer (voir la documentation de SAS). Evidemment si le test F global n'est pas signficatif il est impératif de contrôles le seuil global lorsque l'on fait des comparaisons multiples et d'utiliser la méthode de Tukey.

Lorsque le test F global est significatif. Le problème de contrôle du seuil global ne se pose plus. Dans ce cas les deux méthodes LSD de Fisher et TSD de Tukey sont toutes les deux

acceptables. Evidemment LSD est plus libérale et permettra de trouver plus de différences significatives. Le choix de la méthode dépend souvent des objectifs de l'expérience et du domaine d'application. Une foule de méthode intermédiaire entre ces deux extrêmes ont été proposées dans la littérature et sont souvent utilisées dans des domaines d'application spécifiques (voir la documentation de GLM dans SAS: on y retrouve entre autres les méthodes de SNK et Duncan). Il existe également d'autres méthodes comme celle de Dunnett où on contrôle le seuil global lorsque l'on compare les $I - 1$ premières moyennes à la dernière qui joue le rôle d'un contrôle ou d'un témoin.