Premières notions de statistique: Analyse de la Variance à deux facteurs

Franck Picard

Licence 3 Biosciences, 2022-2023

Laboratoire Biologie et Modélisation de la Cellule, CNRS ENS-Lyon

franck.picard@ens-lyon.fr

Outline

1. **Introduction à l'ANOVA à plusieurs facteurs**
2. Décomposition des sommes de carré
3. Tests des effets des facteurs
4. Comparaison des traitements
5. Introduction au cas non-orthogonal

Généralisation de l'approche de l'ANOVA à plusieurs facteurs

- Lors d'une expérience, plusieurs facteurs sont sources de variation
- On cherche à attribuer la variabilité observée à différentes caractéristiques connues des données
- Exemple historique: On considère un trait quantitatif qui dépend d'un facteur génétique et d'un facteur environnemental
- Facteur Génotype à 3 modalités (`AA`, `Aa`, `aa`), facteur environnemental à deux modalités (chaud/froid).

La difficulté des approches à plusieurs facteurs est l'identification précise des sources de variations observées, en prenant en compte la possible **confusion d'effets**

Exemple de confusion d'effets

- On considère des poules de génotype `AA` et aa et on mesure le poids des oeufs en fonction de deux conditions de température chaud-froid.
- On sait que le génotype n'a pas d'effet sur le poids des oeufs
- On sait que la température a un effet sur le poids des oeufs
- On a soumis tous les individus de génotype `AA` à des témpératures fortes et tous les individus aa à des températures faibles
- On conclut que le génotype a un effet sur le poids des oeufs

C'est la construction du plan d'expérience qui garantira l'interprétabilité des résultats et la capacité du modèle à identifier les sources de variabilités pertinentes

Notations

- On considère un modèle à deux facteurs, donc on introduit des indices pour les modalités des deux
- On note y_{ijk} la $k^{\text{ème}}$ mesure du trait pour un individu de génotype $i = 1 \ldots, I$ ayant évolué dans la condition environnementale $j = 1, \ldots, J$
- On définit $\mathbf{y}_{ij}$ le vecteur des observations pour le croisement des modalités (i, j), de taille n_{ij}
- On peut utiliser aussi les notations $\mathbf{y}_i$ le vecteur contenant les observations pour la modalité i du premier facteur, et toutes les modalités du deuxième (idem avec $\mathbf{y}_j$)
- Ces deux vecteurs sont de taille $n_{i+} = \sum_j n_{ij}$ et $n_{+j} = \sum_i n_{ij}$

La table de contingence

- C'est la table des effectifs croisés entre les niveaux des facteurs
- Elle définit la répartition des observations en fonction des niveaux des facteurs
- Cette table de contingence décrit la répartition des individus dans la matrice **X**

Env. / Géno.	AA ($i=1$)	Aa ($i=2$)	aa ($i=3$)	
Chaud ($j=1$)	$\dots$	n_{ij}	$\dots$	n_{+j}
Froid ($j=2$)	$\dots$		$\dots$	
		n_{i+}		$n = n_{++}$

Intuition autour de la confusion d'effets

- n_{ij}/n_{++} donne la probabilité de tirer une unité expérimentale avec les modalités (i, j)
- $A_i = 1$ si l'individu tiré est de génotype i,
- $B_j = 1$ si l'individu tiré a évolué dans l'environnement j
- $\mathbb{P}(A_i \cap B_j) = n_{ij}/n_{++}$ donne la probabilité de tirer une unité expérimentale avec les modalités (i, j)
- Si A_i et B_j sont indépendants:

$$\mathbb{P}(A_i = 1) = \frac{n_{i+}}{n_{++}} \qquad \mathbb{P}(B_j = 1) = \frac{n_{+j}}{n_{++}}$$

- Dans ce cas

$$n_{ij} = \frac{n_{i+} \times n_{+j}}{n_{++}}$$

Condition d'orthogonalité

- $k = 1, \dots n_{ij}$ est l'indice de répétition, n_{i+} et n_{+j} sont le nombre total d'individus dans la modalité i (j respectivement)
- On appellera **plan équilibré** un plan d'expérience tel que n_{ij} est constante
- On appellera un **plan orthogonal** un plan d'expérience tel que:

$$n_{ij} = \frac{n_{i+} \times n_{+j}}{n_{++}}$$

Lorsque le plan d'expérience sera orthogonal, le modèle permettra de séparer les sources de variabilités et donc de quantifier **séparemment** les contributions de chacun des facteurs

Exemple: rendement des vaches laitières

- Quelle est l'influence de l'alimentation sur le rendement de production en lait de vaches
- On prend en compte l'influence du génotype des animaux

Génotype	Paille	Foin	Herbe	Ensilage	Génotype	Paille	Foin	Herbe	Ensilage
A	8	12	12	14	B	8	10	11	17
A	10	13	10	17	B	9	12	9	19
A	11	11	13	13	B	8	10	11	17
A	10	14	12	14	B	10	11	11	16
A	7	10	14	17	B	9	7	12	21

- On vérifie que le plan d'expérience est équilibré $n_{ij} = 5$

```
> table(A$Genotype,A$Aliment)

        Ensilage Foin Herbe Paille
  A        5    5     5      5
  B        5    5     5      5
```

Extension du modèle pour plusieurs facteurs

- On souhaite expliquer les variations d'un trait en fonction de deux facteurs
- On note $A_i = 1$ si le facteur A est de modalité i, $B_j = 1$ si le facteur B est de modalité j
- Pour chaque combinaison des facteur A et B on a des répétitions $k = 1, \ldots, n_{ij}$.
- On note y_{ijk} la $k^{\text{ème}}$ mesure du trait pour un individu de génotype $i = 1 \ldots, I$ ayant évolué dans la condition environnementale $j = 1, \ldots, J$
- On modélise y_{ijk} par Y_{ijk} et on suppose que l'espérance des observations dépend à la fois du facteur A et du facteur B

$$\mathbb{E}(Y_{ijk} | A_i = 1, B_j = 1) = \mu_{ij}$$

Extension du modèle pour plusieurs facteurs

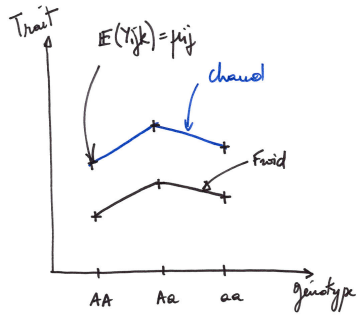
- On note y_{ijk} la $k^{\text{ème}}$ mesure du trait pour un individu de génotype $i = 1 \ldots, I$ ayant évolué dans la condition environnementale $j = 1, \ldots, J$
- On modélise y_{ijk} par Y_{ijk} et on suppose que l'espérance des observations dépend à la fois du facteur A et du facteur B

$$\mathbb{E}(Y_{ijk}|A_i = 1, B_j = 1) = \mu_{ij}$$

- Par exemple $\mu_{\texttt{AA},\text{Chaud}}$ $(i = 1, j = 1)$ est la moyenne du trait des individus qui sont de génotype `AA` et qui sont dans la condition environnementale "Chaud"
- Dans la suite, on ne précisera plus le conditionnement par rapport à A et B (implicite en ANOVA)

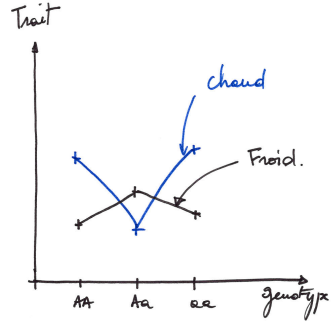
On souhaiterait décomposer ce signal de manière plus fine, pour identifier la part due au génotype et la part due à l'environnement

Effet conjoint de plusieurs facteurs sur une réponse



La Réponse à une différence de Température ne dépend pas du génotype

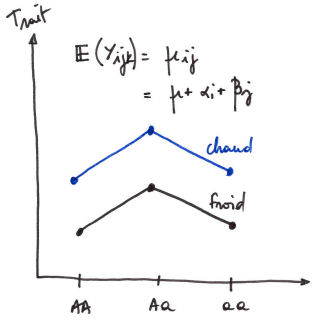
↳ PAS INTERACTION



La Réponse à une différence de Température dépend du Génotype

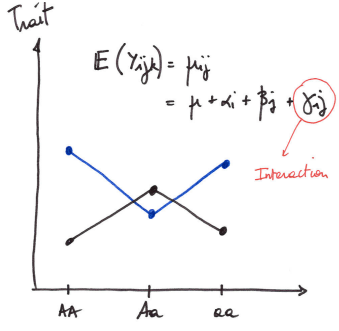
↳ INTERACTION

La notion d'interaction



↳ Le signal peut être modélisé par une somme d'effets

→ MODÈLE ADDITIF



↳ Présence d'INTERACTION entre facteurs

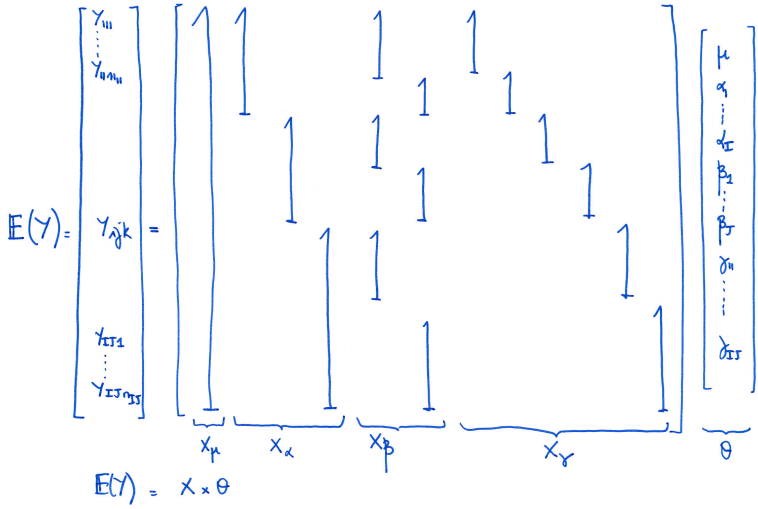
→ γ_{ij} modélise l'interaction du génotype i et de la condit° j sur Y.

Le modèle avec interaction

- Le modèle s'écrit:
$$\mathbb{E}(Y_{ijk}) = \mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$
- α_i et β_j sont les effets **principaux**: c'est l'effet du génotype (marginal, quelque soit la condition environnementale), et l'effet de l'environnement (quelque soit le génotype)
- γ_{ij} est un terme d'**interaction** qui modélise **l'effet conjoint** du génotype et de l'environnement

Le terme d'interaction quantifie l'effet conjoint de plusieurs facteurs sur la réponse

Ecriture matricielle du modèle



Il y a toujours de l'interaction

- L'interaction est une **notion statistique** permettant d'expliquer les effets conjoints de plusieurs facteurs
- C'est la part de variabilité qui ne peut pas être expliquée séparément par chaque facteur et qui n'est pas dans la résiduelle
- Elle est en général présente mais **il n'est pas toujours possible de l'estimer** (dépend de la présence de répétitions)
- Dans l'interprétation des résultats on s'intéressera d'abord aux termes d'interactions pour savoir s'ils sont significatifs

Si le terme d'interaction est très fort et que le plan d'expérience n'est pas orthogonal, alors on ne pourra pas interpréter les effets principaux (non distinguables)

Outline

1. Introduction à l'ANOVA à plusieurs facteurs
2. **Décomposition des sommes de carré**
3. Tests des effets des facteurs
4. Comparaison des traitements
5. Introduction au cas non-orthogonal

Retour sur le théorème de Pythagore

- La stratégie est la même que dans le cas de l'ANOVA à un facteur: on décompose la somme des carrés totale

$$\mathrm{SCT}(\mathbf{Y}) = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}}(Y_{ijk} - Y_{\bullet\bullet\bullet})^2$$

- On cherche à décomposer cette somme avec:

$$\mathrm{SCT}(\mathbf{Y}) = \mathrm{SCM}(\mathbf{Y}, \mathbf{X}) + \mathrm{SCR}(\mathbf{Y}, \mathbf{X})$$

- Pour cela on considère le modèle: $\mathbb{E}(Y_{ijk}) = \mu_{ij}$, avec

$$\widehat{\mu}_{ij} = Y_{ij\bullet}$$

Décomposition des sommes de carré

- La somme des carrés totale se décompose donc:

$$\mathrm{SCT}(\mathbf{Y}) = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}}(Y_{ij\bullet} - Y_{\bullet\bullet\bullet})^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}}(Y_{ijk} - Y_{ij\bullet})^2$$

- La somme des carrés du modèle quantifie l'information apportée par un modèle qui considère que $\mathbb{E}(Y_{ijk}) = \mu_{ij}$

$$\mathrm{SCM}(\mathbf{Y}, \mathbf{X}) = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}}(Y_{ij\bullet} - Y_{\bullet\bullet\bullet})^2$$

- La somme des carrés résiduelle quantifie l'écart des observations au modèle qui considère que $\mathbb{E}(Y_{ijk}) = \mu_{ij}$

$$\mathrm{SCR}(\mathbf{Y}, \mathbf{X}) = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}}(Y_{ijk} - Y_{ij\bullet})^2$$

Ecriture du modèle

- On note Y_{ijk} le rendement d'une vache k ayant reçu l'aliment $i = 1 \dots 4$ à la dose $j = 1 \dots J$, k est l'indice de répétition ($k = 1...5$)
- On note r le nombre de répétitions dans chaque cellule:

$$n_{++} = r \times IJ$$

- On suppose que le rendement moyen dépend de l'aliment et de la dose tel que: $\mathbb{E}(Y_{ijk}) = \mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$
- α_i est l'effet principal de l'aliment sur le rendement
- β_j est l'effet principal de la dose d'aliment sur le rendement
- γ_{ij} est un terme d'interaction qui modélise l'effet conjoint de la dose et de l'aliment
- On suppose que le modèle s'écrit:

$$Y_{ijk} = \mu_{ij} + E_{ijk}, \qquad E_{ijk} \underset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2),\ \sigma^2 \text{constante}$$

Le modèle est-il pertinent-1?

- On s'interroge quant à la pertinence **globale** du modèle
- On teste le modèle nul contre le modèle complet:

$$H_0 : \{\forall (i,j)\ \mu_{ij} = \mu\}$$

- L'espérance des sommes de carrés s'écrit:

$$\begin{aligned}
\mathbb{E}(\mathrm{SCM}(\mathbf{Y},\mathbf{X})) &= \sum_{i,j} n_{ij}\mathbb{E}(Y_{ij\bullet} - Y_{\bullet\bullet\bullet})^2 \\
&= (IJ-1)\sigma^2 + \sum_{ij} n_{ij}(\mu_{ij}-\mu)^2 \\
\frac{\mathrm{SCM}(\mathbf{Y},\mathbf{X})}{IJ-1} &\underset{H_0}{\sim} \sigma^2\chi^2(IJ-1) \\
\frac{\mathrm{SCR}(\mathbf{Y},\mathbf{X})}{IJ(r-1)} &\underset{H_0}{\sim} \sigma^2\chi^2(IJ(r-1))
\end{aligned}$$

Le modèle est-il pertinent-2?

- Le premier test est celui du modèle "global":

$$F = \frac{\text{SCM}(\mathbf{Y}, \mathbf{X})/(IJ-1)}{\text{SCR}(\mathbf{Y}, \mathbf{X})/(IJ(r-1))} \underset{H_0}{\sim} \mathcal{F}(IJ-1, IJ(r-1))$$

- Sur l'exemple des vaches:

```
Residual standard error: 3.337 on 32 degrees of freedom
Multiple R-squared:  0.334,Adjusted R-squared:  0.1883
F-statistic: 2.292 on 7 and 32 DF,  p-value: 0.05182
```

- le modèle est faiblement pertinent mais explique 34% de la variabilité (18% si on corrige par le nombre de paramètres)

Outline

1. Introduction à l'ANOVA à plusieurs facteurs
2. Décomposition des sommes de carré
3. **Tests des effets des facteurs**
4. Comparaison des traitements
5. Introduction au cas non-orthogonal

La décomposition de la somme des carrés du modèle

- Après avoir étudié l'apport global du modèle μ_{ij} on peut se poser des questions plus fines sur la structuration du modèle qui est décomposé en:

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

- la variabilité des observations Y_{ijk} autour de $Y_{ij\bullet}$ peut être décomposée en fonction de $Y_{i\bullet\bullet}$ et $Y_{\bullet j\bullet}$:

$$\begin{aligned} \text{SCM}(\mathbf{Y}, \mathbf{X}) &= \sum_{ijk} (Y_{i\bullet\bullet} - Y_{\bullet\bullet\bullet})^2 + \sum_{ijk} (Y_{\bullet j\bullet} - Y_{\bullet\bullet\bullet})^2 \\ &+ \sum_{ijk} (Y_{ij\bullet} - Y_{i\bullet\bullet} - Y_{\bullet j\bullet} + Y_{\bullet\bullet\bullet})^2 \end{aligned}$$

Cette décomposition est **unique** si le dispositif est **orthogonal**

Hypothèses pour tester les effets des facteurs

- Plus synthétiquement:

$$\text{SCM}(\mathbf{Y}, \mathbf{X}) = \text{SCM}(\mathbf{Y}, \mathbf{X}_\alpha) + \text{SCM}(\mathbf{Y}, \mathbf{X}_\beta) + \text{SCM}(\mathbf{Y}, \mathbf{X}_\gamma)$$

- Chaque somme de carrés quantifie la contribution de chaque facteur à la variabilité expliquée par le modèle
- Dans le cas d'un plan orthogonal, on peut donc quantifier séparément l'apport de chaque facteur
- Les trois familles d'hythèses sont donc:

$$\begin{aligned} H_0^\gamma &: \{\gamma_{11} = \ldots = \gamma_{IJ} = 0\} \text{ Pas d'interaction} \\ H_0^\alpha &: \{\alpha_1 = \ldots = \alpha_I = 0\} \text{ Pas d'effet du facteur A} \\ H_0^\beta &: \{\beta_1 = \ldots = \beta_J = 0\} \text{ Pas d'effet du facteur B} \end{aligned}$$

Tests les effets des facteurs (cas orthogonal)

- Plus synthétiquement:

$$\begin{aligned}
\mathrm{SCM}(\mathbf{Y}, \mathbf{X}_\alpha)/(I-1) &\underset{H_0^\alpha}{\sim} \sigma^2\chi^2(I-1) \\
\mathrm{SCM}(\mathbf{Y}, \mathbf{X}_\beta)/(J-1) &\underset{H_0^\beta}{\sim} \sigma^2\chi^2(J-1) \\
\mathrm{SCM}(\mathbf{Y}, \mathbf{X}_\gamma)/(I-1)(J-1) &\underset{H_0^\gamma}{\sim} \sigma^2\chi^2((I-1)(J-1))
\end{aligned}$$

- On peut tester chaque somme de carré séparément à la somme des carrés résiduelle:

$$\begin{aligned}
F_\alpha &= \frac{\mathrm{SCM}(\mathbf{Y}, \mathbf{X}_\alpha)/(I-1)}{\mathrm{SCR}(\mathbf{Y}, \mathbf{X})/(IJ(r-1))} \\
F_\beta &= \frac{\mathrm{SCM}(\mathbf{Y}, \mathbf{X}_\beta)/(J-1)}{\mathrm{SCR}(\mathbf{Y}, \mathbf{X})/(IJ(r-1))} \\
F_\gamma &= \frac{\mathrm{SCM}(\mathbf{Y}, \mathbf{X}_\alpha)/(I-1)(J-1)}{\mathrm{SCR}(\mathbf{Y}, \mathbf{X})/(IJ(r-1))}
\end{aligned}$$

La table d'ANOVA pour plusieurs effets

Source	Sum of Sq.	dF	Mean Squares	F-stat	Pv
Factor A	SCM_A	I-1	SCM_A / (I-1)	MSM_A/MSR	.
Factor B	SCM_B	J-1	SCM_B / (J-1)	MSM_B/MSR	.
Factor AB	SCM_{AB}	(I-1)(J-1)	SCM / (I-1)(J-1)	MSM_{AB}/MSR	.
Residual	SCR	IJ(r-1)	SCR / (IJr-1)		
Total	SCT	rIJ-1			

```
Response: rdt
             Df Sum Sq Mean Sq F value  Pr(>F)
Dose          1   36.1  36.100  3.2413 0.08124 .
Aliment       3  119.3  39.767  3.5705 0.02466 *
Dose:Aliment  3   23.3   7.767  0.6973 0.56054
Residuals    32  356.4  11.137
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Facteurs multiples

- Dans le cas de plus de deux facteurs, la combinatoire explose pour les interactions d'ordre multiple
- On fait en général l'hypothèse que l'intensité des effets décroit avec l'ordre de l'interaction
- Dans un modèle avec 4 facteurs, on étudiera en général les effets principaux et les interactions d'ordre 2 voire 3
- Pour des modèles avec beaucoup de facteurs, il est crucial de bien comprendre la signification de chaque paramètre

Comment identifier les facteurs pertinents ?

Exemple en génomique: force des origines de réplication

- On étudie la position des origines de réplication le long du génome humain. On sait mesurer la *force* d'une origine, i.e. le nombre de cellules qui utilisent une origine particulière
- On cherche à déterminer quels facteurs génomiques influencent la force des origines en fonction de covariables:

Code	Facteur
`timing.class.6`	timing de réplication (6 modalités)
`cgi`	Présence / Absence d'un ilôt CpG
`gquad.1.7`	Présence / Absence d'un G-quadruplex
`H2az`	Présence / Absence du variant H2az de l'histone H2A
`H4k20me1`	Présence / Absence d'une histone H4 1-methylée k20
`H3k27me3`	Présence / Absence d'une histone H3 3-methylée k27
`H3k9me3`	Présence / Absence d'une histone H3 3-methylée k9
`H3k9ac`	Présence / Absence d'une histone H3 acethylée k9
`H3k4me3`	Présence / Absence d'une histone H3 3-methylée k4

Comment identifier les effets pertinents ?

	Sum Sq	Df	F value	Pr(>F)
timing.class.6	7.24	6	573.94	0.0000
cgi	0.54	1	257.07	0.0000
gquad.1.7	0.31	1	146.61	0.0000
timing.class.6:cgi	0.50	5	47.52	0.0000
H4k20me1:H3k27me3	0.10	1	47.30	0.0000
gquad.1.7:H4k20me1	0.07	1	34.95	0.0000
H3k27me3:H3k9me3	0.07	1	33.16	0.0000
timing.class.6:gquad.1.7	0.23	5	21.87	0.0000
H3k9me3:H3k9ac	0.04	1	20.86	0.0000
timing.class.6:H3k9me3	0.20	5	19.11	0.0000
H3k9me3	0.03	1	15.63	0.0001
H3k9ac:H3k4me3	0.03	1	14.66	0.0001
H3k4me3	0.03	1	14.51	0.0001
H2az	0.03	1	14.30	0.0002
H4k20me1	0.03	1	14.08	0.0002
H2az:H3k9ac	0.03	1	12.26	0.0005
⋮				
Residuals	77.61	36914		$R^2 = 0.75$

Comment identifier les effets pertinents ?

- Les Pvalues peuvent être utiles: elles quantifient la significativité de chaque test
- Lorsque beaucoup d'observations sont disponibles, les tests deviennent puissants, et identifieront des interactions faibles mais significatives
- Les Pvalues peuvent être plus petites que la précision machine ($\sim 10^{-16}$): comment les comparer ?
- Les statistiques de Fisher peuvent être utilisées pour comparer les effets des facteurs et de leurs interactions

	Sum Sq	Df	F value	Pr(>F)
timing.class.6	7.24	6	573.94	0.0000
cgi	0.54	1	257.07	0.0000
gquad.1.7	0.31	1	146.61	0.0000
timing.class.6:cgi	0.50	5	47.52	0.0000
H4k20me1:H3k27me3	0.10	1	47.30	0.0000
⋮				
Residuals	77.61	36914		

Prendre en compte la multiplicité des tests

- Face à un grand nombre de facteurs et de termes d'interactions, se pose la question des tests multiples
- On peut ajuster les Pvalues des tests de Fisher

	Sum Sq	Df	F value	Ajusted Pr(>F)
timing.class.6	7.24	6	573.94	0.0000
cgi	0.54	1	257.07	0.0000
gquad.1.7	0.31	1	146.61	0.0000
timing.class.6:cgi	0.50	5	47.52	0.0000
H4k20me1:H3k27me3	0.10	1	47.30	0.0000
gquad.1.7:H4k20me1	0.07	1	34.95	0.0000
H3k27me3:H3k9me3	0.07	1	33.16	0.0000
timing.class.6:gquad.1.7	0.23	5	21.87	0.0000
H3k9me3:H3k9ac	0.04	1	20.86	0.0002
timing.class.6:H3k9me3	0.20	5	19.11	0.0000
H3k9me3	0.03	1	15.63	0.0035
H3k9ac:H3k4me3	0.03	1	14.66	0.0058
⋮				
Residuals	77.61	36914		

Outline

1. Introduction à l'ANOVA à plusieurs facteurs
2. Décomposition des sommes de carré
3. Tests des effets des facteurs
4. **Comparaison des traitements**
5. Introduction au cas non-orthogonal

Estimation des paramètres

- La décomposition du modèle μ_{ij} permet de mieux interpréter les différents effets mais pose un problème d'identifiabilité
- Comme dans le cas du modèle à un facteur, on pose certaines contraintes pour l'estimation des paramètres par le critère des moindre-carrés:

$$d^2(\mathbf{Y}; \mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{ijk} (Y_{ijk} - [\mu + \alpha_i + \beta_j + \gamma_{ij}])^2$$

- On pose les contraintes usuelles

$$\alpha_I = 0 \ \beta_J = 0 \ \forall i, \gamma_{iJ} = 0, \ \ \forall j, \gamma_{Ij} = 0$$

- Il faut utiliser la contrainte $\sum_i n_{i+}\alpha_i = 0$, $\sum_j n_{+j}\beta_j = 0$, $\forall j$, $\sum_i n_{ij}\gamma_{ij} = 0$, $\forall i$, $\sum_j n_{ij}\gamma_{ij} = 0$, pour retrouver les estimateurs naturels

Comparaison des traitements entre eux

- On peut s'interroger sur la comparaison des effets des traitements (exemple, trouver le génotype associé au meilleur rendement):

$$H_0^{\alpha} : \{\alpha_i = \alpha_{i'}\} \qquad H_0^{\beta} : \{\beta_j = \beta_{j'}\}$$

- Il faut être prudent sur la confusion d'effet
- Si le plan n'est pas orthogonal, $\widehat{\alpha}_i - \widehat{\alpha}_{i'}$ contient de l'information liée aux autres facteurs (différence due à l'environnement et pas au génotype par exemple)
- Si on ne s'intéresse qu'au premier facteur, on construit une moyenne **ajustée**

Test sur les moyennes ajustées

- On **s'ajuste** sur les niveaux de l'autre facteur

$$\widetilde{\mu}_i = \frac{1}{J}\sum_{j=1}^{J} \widehat{\mu}_{ij} = \widehat{\mu} + \widehat{\alpha}_i + \sum_j \left(\widehat{\beta}_j + \widehat{\gamma}_{ij}\right)$$

- On teste $H_0^\alpha : \{\alpha_i = \alpha_{i'}\}$ à l'aide d'une nouvelle statistique fondée sur:

$$\widetilde{\mu}_i - \widetilde{\mu}_{i'} = \widehat{\alpha}_i - \widehat{\alpha}_{i'} + \frac{1}{J}\sum_j \left(\widehat{\gamma}_{ij} - \widehat{\gamma}_{i'j}\right)$$

- Si le plan n'est pas orthogonal et que l'interaction est significative, cette différence est **corrigée** de l'effet de l'interaction
- On n'attribuera pas de différence de traitement si elle est due à l'autre facteur confondu

Outline

1. Introduction à l'ANOVA à plusieurs facteurs
2. Décomposition des sommes de carré
3. Tests des effets des facteurs
4. Comparaison des traitements
5. **Introduction au cas non-orthogonal**

Tout s'écroule !

- C'est le cas le plus rencontré en pratique
- Ces plans d'expérience ne permettent pas de distinguer les effets des différents facteurs de manière unique
- Lorsque l'on effectue un test sur les α_i, on ne peut pas s'affranchir des autres effets
- La décomposition de la somme des carrés du modèle n'est plus unique, on a donc plusieurs façons de la décomposer
- Plusieurs tests possibles (sommes de type I, II, III)