Démarche statistique

Session 8 - Planification d'expériences

David Causeur
Institut Agro Rennes Angers
IRMAR UMR 6625 CNRS

19 janvier, 2025

Pourquoi planifier le recueil de données ?

Plan d'expériences

Objectif général : Organiser le recueil de données, pour estimer **au mieux** les paramètres d'un modèle (linéaire).

- Plus le nombre n d'individus dans l'échantillon est grand, meilleure est la précision de l'estimation.
- A nombre d'individus fixé, la **précision** de l'estimation dépend du choix de ces individus selon leurs valeurs des variables explicatives.

Fisher, R. A. (1935) The Design of Experiments. (9th ed.) : principes fondamentaux de la planification expérimentale illustrés par le défi du *thé au lait* (The Lady tasting tea).

Illustration : On cherche à estimer un modèle de régression expliquant **la teneur en viande maigre** d'une carcasse de porc à partir d'**épaisseurs de gras** mesurés en différents sites anatomiques. Le budget expérimental permet un échantillon de $n = 100$ carcasses.

Comment choisir ces 100 individus à partir desquels on va estimer le modèle ?

- Au hasard uniforme dans la population ?
- En privilégiant des individus avec des épaisseurs de gras proches de la moyenne ?
- En partageant l'échantillon en 2 : 50 individus avec de faibles épaisseurs de gras et 50 individus avec de fortes épaisseurs de gras ?

Précision de l'estimation des effets

Illustration : **échantillonnage** (choix des individus) en régression linéaire simple

$$\text{Pour } i = 1, \ldots, n, \;\; Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \;\; \varepsilon_i \sim \mathcal{N}(0; \sigma)$$

Estimation par la méthode des moindres carrés : $\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}$

Précision de l'estimation :

$$\text{Var}\left(\hat{\beta}_1\right) \quad = \quad \frac{\sigma^2}{n} \frac{1}{s_x^2}$$

La précision de l'estimation d'un effet est d'autant meilleure que :

- σ est proche de 0 : le modèle approche bien la relation entre Y et x
- n est grand : l'échantillon est de grande taille
- s_x^2 **est grand** : les valeurs de x sont dispersées

En pratique, on choisit la moitié de l'échantillon parmi les individus ayant une valeur faible de x, l'autre moitié parmi les individus ayant une valeur élevée de x

Confusions d'effets

Illustration : On cherche à optimiser l'appréciation sensorielle (Y) d'un biscuit, dont la recette dépend de :

- la quantité de lait dans la pâte ($x_1 \in \{B, H\}$)
- la température de cuisson ($x_2 \in \{B, H\}$)
- la quantité de sucre ($x_3 \in \{B, H\}$)

Plan d'expériences (liste d'*essais*) complet 2^3

Essai	x_1	x_2	x_3
1	H	H	H
2	H	H	B
3	H	B	H
4	H	B	B
5	B	H	H
6	B	H	B
7	B	B	H
8	B	B	B

Comment répartir au mieux ces essais entre deux équipes ?

Principes généraux des plans pour facteurs à deux modalités

Estimation d'une différence de moyennes : analyse de la variance à un facteur

Illustration : modèle d'analyse de la variance à un facteur (à 2 modalités)

$$\text{Pour } i = 1, 2,\ j = 1, \ldots, n_i,\ Y_{ij} = \mu + \alpha_i + \varepsilon_{ij},\ \varepsilon_{ij} \sim \mathcal{N}(0; \sigma)$$

Estimation d'une différence de moyennes (entre les groupes 1 et 2) :

$$\hat{\alpha}_2 = \bar{Y}_2 - \bar{Y}_1,\ \text{Var}\left(\hat{\alpha}_2\right) = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right).$$

Impact de la répartition des individus dans les deux groupes :

- $n = n_1 + n_2$
- Soit $f = \frac{n_1}{n}$, alors $0 \leq f \leq 1$ et $n_1 = fn,\ n_2 = (1 - f)n$.

$$\begin{aligned} \text{Var}\left(\hat{\alpha}_2\right) &= \frac{\sigma^2}{n} \left(\frac{1}{f} + \frac{1}{1-f} \right), \\ &= \frac{\sigma^2}{n} \left(\frac{1}{f(1-f)} \right), \text{ minimal pour } f = \frac{1}{2} \\ &\geq \frac{4\sigma^2}{n} \end{aligned}$$

Effet d'une variable catégorielle : le dispositif optimal est **équilibré** ($n_1 = n_2$)

Confusion entre deux facteurs (à deux modalités)

Analyse de la variance à deux facteurs, x_1 et x_2, chacun à deux modalités (B et H)

Pour $i = 1, 2,\ j = 1, 2,\ k = 1, \ldots, n_{ij},\ Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk},\ \varepsilon_{ijk} \sim \mathcal{N}(0; \sigma)$

Dispositif expérimental :

- Pour chaque facteur, $n/2$ individus avec la modalité H
- f, proportion des individus ayant la modalité H pour x_1 parmi ceux ayant la modalité H pour x_2

	x_2		
x_1	B	H	Total
B	$\frac{n}{2}$ f	$\frac{n}{2}(1-f)$	$\frac{n}{2}$
H	$\frac{n}{2}(1-f)$	$\frac{n}{2}f$	$\frac{n}{2}$
Total	$\frac{n}{2}$	$\frac{n}{2}$	n

Précision de l'estimation par la méthode des moindres carrés :

$$\mathrm{Var}\left(\hat{\alpha}_2\right) = \mathrm{Var}\left(\hat{\beta}_2\right) = \frac{4\sigma^2}{n}\,\frac{1}{1-(1-2f)^2},$$

- Précision optimale si $f = \frac{1}{2}$ (plan complet et équilibré)
- Précision dégradée si $f \approx 0$ ou $f \approx 1$ (confusion entre x_1 et x_2)

Plan complet pour facteurs à deux modalités

Plan complet pour p facteurs à deux modalités: 2^p combinaisons possibles des modalités des p facteurs

Illustration : matrice des essais (en ordre aléatoire) pour 3 facteurs

```
plan3 <- FrF2::FrF2(nruns=8,nfactors=3,factor.names=paste0("F",1:3),
                    replications=1)
print(plan3)
```

```
  F1 F2 F3
1 -1 -1  1
2  1 -1 -1
3  1  1 -1
4 -1  1  1
5 -1 -1 -1
6  1  1  1
7 -1  1 -1
8  1 -1  1
class=design, type= full factorial
```

Remarques :

- la non-confusion complète entre deux facteurs F_i et F_j se traduit par $F_i.F_j=0$, où $F_i.F_j$ est le produit scalaire des vecteurs colonnes, soit la somme des produits des coordonnées.
- on cherche en pratique à construire des plans qui garantissent au moins la non-confusion entre les **effets principaux** des facteurs

Codages binaires des modalités

Illustration: analyse de la variance à deux facteurs à deux modalités

- **Contraintes sur les paramètres par défaut dans la fonction lm :**
  - $\alpha_1 = 0$, $\beta_1 = 0$,
  - $(\alpha\beta)_{11} = 0$, $(\alpha\beta)_{12} = 0$ et $(\alpha\beta)_{21} = 0$

$$
\begin{aligned}
Y_{11} &= 1 \times \mu + 0 \times \alpha_2 + 0 \times \beta_2 + 0 \times (\alpha\beta)_{22} + \varepsilon_{11} \\
Y_{12} &= 1 \times \mu + 0 \times \alpha_2 + 1 \times \beta_2 + 0 \times (\alpha\beta)_{22} + \varepsilon_{12} \\
Y_{21} &= 1 \times \mu + 1 \times \alpha_2 + 0 \times \beta_2 + 0 \times (\alpha\beta)_{22} + \varepsilon_{21} \\
Y_{22} &= 1 \times \mu + 1 \times \alpha_2 + 1 \times \beta_2 + 1 \times (\alpha\beta)_{22} + \varepsilon_{22}
\end{aligned}
$$

- **Contraintes sur les paramètres dans la fonction LinearModel de FactoMineR :**
  - $\alpha_1 + \alpha_2 = 0$, $\beta_1 + \beta_2 = 0$,
  - $(\alpha\beta)_{11} + (\alpha\beta)_{12} = 0$, $(\alpha\beta)_{21} + (\alpha\beta)_{22} = 0$, $(\alpha\beta)_{11} + (\alpha\beta)_{21} = 0$ et $(\alpha\beta)_{12} + (\alpha\beta)_{22} = 0$

$$
\begin{aligned}
Y_{11} &= 1 \times \mu - 1 \times \alpha_2 - 1 \times \beta_2 + 1 \times (\alpha\beta)_{22} + \varepsilon_{11} \\
Y_{12} &= 1 \times \mu - 1 \times \alpha_2 + 1 \times \beta_2 - 1 \times (\alpha\beta)_{22} + \varepsilon_{12} \\
Y_{21} &= 1 \times \mu + 1 \times \alpha_2 - 1 \times \beta_2 - 1 \times (\alpha\beta)_{22} + \varepsilon_{21} \\
Y_{22} &= 1 \times \mu + 1 \times \alpha_2 + 1 \times \beta_2 + 1 \times (\alpha\beta)_{22} + \varepsilon_{22}
\end{aligned}
$$

Matrice des effets pour un modèle avec interaction

Matrice des essais et matrice des effets associée

	Matrice des essais	
Essai	F_1	F_2
1	H	H
2	H	B
3	B	H
4	B	B

Matrice des effets			
I	F_1	F_2	$F_1 : F_2$
+1	+1	+1	+1
+1	+1	-1	-1
+1	-1	+1	-1
+1	-1	-1	+1

Remarques

- Le codage binaire de l'effet d'interaction $F_1 : F_2$ s'obtient à partir du produit terme à terme des vecteurs colonnes associés à F_1 et F_2
- La matrice des effets permet de déterminer les séquences d'essais possibles garantissant :
  - l'équilibre du plan (autant de +1 que de -1)
  - la non-confusion avec les effets principaux (produit nul avec les vecteurs associés aux effets principaux)

Plans fractionnaires

Dispositifs expérimentaux incomplets

Illustration : On cherche à optimiser l'appréciation sensorielle d'un biscuit, dont la recette dépend de :

- la quantité de lait dans la pâte ($x_1 \in \{B, H\}$)
- la température de cuisson ($x_2 \in \{B, H\}$)
- la quantité de sucre ($x_3 \in \{B, H\}$)

Les mesures sont réparties entre deux équipes d'opératrices.teurs ($x_4 \in \{1, 2\}$)

Plan d'expériences (incomplet) : plan fractionnaire 2^{4-1}

Essai	x_1	x_2	x_3	x_4
1	H	H	H	?
2	H	H	B	?
3	H	B	H	?
4	H	B	B	?
5	B	H	H	?
6	B	H	B	?
7	B	B	H	?
8	B	B	B	?

Comment répartir au mieux ces essais entre deux équipes ?

Construction d'un plan fractionnaire optimal

Illustration : plan 2^{4-1}

- Matrice des effets du plan complet 2^3

I	F_1	F_2	F_3	F_1 F_2	F_1 F_3	F_2 F_3	F_1 F_2 F_3
+1	+1	+1	+1	+1	+1	+1	+1
+1	+1	+1	-1	+1	-1	-1	-1
+1	+1	-1	+1	-1	+1	-1	-1
+1	+1	-1	-1	-1	-1	+1	+1
+1	-1	+1	+1	-1	-1	+1	-1
+1	-1	+1	-1	-1	+1	-1	+1
+1	-1	-1	+1	+1	-1	-1	+1
+1	-1	-1	-1	+1	+1	+1	-1

- Liste d'essais pour le facteur supplémentaire F_4 : $F_4 = F_1$ F_2 F_3

Essai	x_1	x_2	x_3	x_4
1	H	H	H	Equipe 2
2	H	H	B	Equipe 1
3	H	B	H	Equipe 1
4	H	B	B	Equipe 2
5	B	H	H	Equipe 1
6	B	H	B	Equipe 2
7	B	B	H	Equipe 2
8	B	B	B	Equipe 1

Niveau global de confusion d'un plan d'expériences : résolution

Illustration : plan 2^{4-1}

- **Générateur d'alias** : $F_4 = F_1\ F_2\ F_3 \Rightarrow I = F_1\ F_2\ F_3\ F_4$
- **Le générateur d'alias est utile pour recenser les confusions** :
  - Effet confondu avec F_1 : $F_1 = F_2\ F_3\ F_4$
  - Effet confondu avec $F_1\ F_3$: $F_1\ F_3 = F_2\ F_4$
- **Mesure du niveau global de confusion** : la résolution du plan est la longueur du plus petit générateur d'alias
  - la résolution du plan 2^{4-1} est donc 4
  - les effets principaux sont confondus avec des interactions d'ordre 3
  - les effets d'interaction d'ordre 2 sont confondus entre eux : pour en estimer certains, d'autres doivent être supposés inexistants
- **Résolutions dégradées** :
  - Résolution 3 : les effets principaux sont confondus avec les effets d'interaction d'ordre 2
  - Résolution 2 : les effets principaux sont confondus entre eux

Plan 2^{4-1} dans R

Illustration : matrice des essais pour 4 facteurs avec $n = 8$

```
plan4 <- FrF2::FrF2(nruns=8,nfactors=4,factor.names=paste0("F",1:4),
                    replications=1)
plan4
```

```
  F1 F2 F3 F4
1  1 -1 -1  1
2 -1  1 -1  1
3 -1 -1 -1 -1
4 -1  1  1 -1
5  1  1  1  1
6  1  1 -1 -1
7 -1 -1  1  1
8  1 -1  1 -1
class=design, type= FrF2
```

```
design.info(plan4)$catlg.entry
```

```
Design:  4-1.1
   8  runs,  4  factors,
   Resolution  IV
   Generating columns:  7
   WLP (3plus):  0 1 0 0 0 ,  0  clear 2fis
```

```
design.info(plan4)$aliased$fi2
```

```
[1] "AB=CD" "AC=BD" "AD=BC"
```

Construction d'un plan fractionnaire avec plusieurs générateurs d'alias

Illustration : plan 2^{5-2}

- Matrice des effets du plan complet 2^3

I	F_1	F_2	F_3	F_1 F_2	F_1 F_3	F_2 F_3	F_1 F_2 F_3
+1	+1	+1	+1	+1	+1	+1	+1
+1	+1	+1	-1	+1	-1	-1	-1
+1	+1	-1	+1	-1	+1	-1	-1
+1	+1	-1	-1	-1	-1	+1	+1
+1	-1	+1	+1	-1	-1	+1	-1
+1	-1	+1	-1	-1	+1	-1	+1
+1	-1	-1	+1	+1	-1	-1	+1
+1	-1	-1	-1	+1	+1	+1	-1

- Proposition de listes d'essais pour les facteurs supplémentaires F_4 et F_5 :
  - $F_4 = F_1\ F_2\ F_3$
  - $F_5 = F_1\ F_2$
- Générateurs d'alias
  - $I = F_1\ F_2\ F_3\ F_4$
  - $I = F_1\ F_2\ F_5$
  - $I = F_3\ F_4\ F_5$
- Résolution du plan : 3

Comment obtenir une meilleure résolution ?

Plan 2^{5-2} dans R

Illustration : matrice des essais pour 5 facteurs avec $n = 8$

```
plan5 <- FrF2::FrF2(nruns=8,nfactors=5,factor.names=paste0("F",1:5))
plan5
```

```
  F1 F2 F3 F4 F5
1 -1  1  1 -1 -1
2 -1  1 -1 -1  1
3  1  1 -1  1 -1
4 -1 -1  1  1 -1
5  1 -1 -1 -1 -1
6 -1 -1 -1  1  1
7  1 -1  1 -1  1
8  1  1  1  1  1
class=design, type= FrF2
```

```
design.info(plan5)$catlg.entry
```

```
Design:  5-2.1
   8  runs,  5  factors,
   Resolution  III
   Generating columns:  3 5
   WLP (3plus):  2 1 0 0 0 ,  0  clear 2fis
```

```
design.info(plan5)$aliased$main
```

```
[1] "A=BD=CE" "B=AD"    "C=AE"    "D=AB"    "E=AC"
```

Ce qu'il faut retenir

La précision de l'estimation des effets et donc la puissance des tests des effets dépendent du choix des individus constituant l'échantillon

- **Effets de variables explicatives quantitatives** : favoriser la dispersion des valeurs dans l'échantillon
- **Effets de variables explicatives catégorielles** : favoriser une répartition equilibrée des combinaisons de modalités des variables, pour limiter les confusions d'effets
- **Cas de p variables explicatives catégorielles à 2 modalités** :
  - Le plan complet 2^p garantit la non-confusion complète mais est très coûteux si p est grand
  - Les plans fractionnaires 2^{p-k} offrent des garanties d'optimalité mais génèrent des confusions d'effets
  - Les stratégies de construction de plans fractionnaires visent donc à minimiser le niveau global de confusion et à permettre l'estimation des effets les plus importants au regard de la problématique

Perspectives

- Module de plan d'expériences en M1 : généralisation à des plans pour variables quantitatives et catégorielles à plus de deux modalités