Randomisation pour (petits) plans d'expériences

Hervé Monod

INRA - Mathématiques et Informatique Appliquées (MIA)
Jouy-en-Josas, France

Journée "Big ideas for small data"
13 mai 2019, AgroParisTech

Plan

Introduction . . . étymologique [wiktionary]

L'hiver survint avec grande furie,
Monceaux de neige et grands **randons** *de pluie.*

(Jean de La Fontaine, Poésies mêlées, XXVII)

Introduction . . . étymologique [wiktionary]

L'hiver survint avec grande furie,
Monceaux de neige et grands **randons** *de pluie.*
(Jean de La Fontaine, Poésies mêlées, XXVII)

- **randon**: Course impétueuse, afflux impétueux.
- **randir**: Courir, galoper. Apparenté à run ("courir"), en anglais, rennen (id.) en allemand ; du vieux-francique rand ("course").
- **randomiser**: de l'anglais randomize, dérivé de l'adjectif random ("aléatoire"), lui-même issu de l'ancien français randon.
- **randonnée**: (Chasse) Tour, circuit que fait autour du même lieu une bête qui, après avoir été lancée, se fait chasser dans son enceinte avant de l'abandonner. (Par extension) Longue promenade, excursion pouvant durer jusqu'à plusieurs jours et revenant à son point de départ.

Introduction

Chez l'humain, le recours à l'aléatoire est fréquent pour jouer ou aider à la décision

- pile ou face, loterie, dés, roulette, cartes, ...
- jouer, parier, sélectionner, organiser, ordonner, arbitrer, ...
- Propriétés recherchées: imprédictibilité, objectivité, transparence, équité
- Intérêt qui est à la source de la théorie des probabilités

Introduction

Aujourd'hui, en statistique

- Randomiser = introduire un élément aléatoire dans un calcul ou dans un raisonnement.
- Utilisé pour
  - échantillonner
  - planifier des expériences
  - conduire des essais cliniques séquentiels et/ou adaptatifs
  - analyser des données (techniques de ré-échantillonnage, tests de permutation, etc.)
- Complémentaire à des étapes devant garantir l'efficacité et la partinence
  - choix des conditions expérimentales
  - identification et prise en compte des sources d'hétérogénéité (blocking, arrière-effets, tendances)
  - recherche d'un plan (initial) optimal

Introduction

Pourquoi randomiser?

- la randomisation a pour but de
  - éviter des choix subjectifs
  - assurer et *faire preuve* d'objectivité et d'équité
  - éliminer des biais
  - valider des modèles ou des procédures de test sous des hypothèses faibles
- cas des très petites expériences
  - risque de tirages inappropriés
  - besoin de méthodes d'inférence robustes

Plan

Plan

Introduction

Contexte: expériences comparatives

- plans factoriels (Fisher et Yates, 1925; agriculture) ou plans pour surfaces de réponse (Box, Taguchi, 1950; chimie)
- **R**eplication, **R**andomisation, **B**locking
  - variables d'intérêt: facteurs "traitements"
  - sources de nuisance connues: facteurs "blocs"
  - sources de nuisance inconnues: randomisation

Méthodes de randomisation des plans en blocs généralisés

Randomisation: à partir d'un plan initial

- Plan complétement aléatoire
  - permutation aléatoire quelconque des unités
- Plan en blocs
  - permutation aléatoire des blocs
  - permutations aléatoires indépendantes des unités dans les blocs
- Plan en lignes-colonnes
  - permutation aléatoire des lignes
  - permutation aléatoire des colonnes

Méthodes de randomisation des plans en blocs généralisés

Randomisation: à partir d'un plan initial

- Plan complétement aléatoire
  - permutation aléatoire quelconque des unités
- Plan en blocs
  - permutation aléatoire des blocs
  - permutations aléatoires indépendantes des unités dans les blocs
- Plan en lignes-colonnes
  - permutation aléatoire des lignes
  - permutation aléatoire des colonnes

$\Longrightarrow$ indépendante du plan initial

$\Longrightarrow$ cohérente avec la structure en blocs choisie

$\Longrightarrow$ détermine le modèle utilisé pour analyser les résultats

$\Longrightarrow$ se généralise à des structures en blocs très générales (Bailey, 1991)

Example 1: plan en blocs

En entrée:

- Facteurs: variété ($m_V = 6$), bloc ($m_B = 4$)
- $N = m_V \times m_B = 24$ unités
- Structure des unités: $\sim$ bloc / parcelle
- Modèle: Y $\sim$ bloc + variété $+\varepsilon$

Construction: plan initial : toutes les combinaisons $V \times B$

A	B	C	D	E	F
A	B	C	D	E	F
A	B	C	D	E	F
A	B	C	D	E	F

+ randomisation $\Rightarrow$

E	C	A	F	D	B
B	E	D	C	F	A
D	E	C	B	F	A
B	D	F	C	E	A

Example 2: plan en lignes - colonnes

En entrée:

- Facteurs: produit (5), juge (5), période (5)
- $N = 25$. Structure des unités: $\sim$ juge * période
- Modèle: Y $\sim$ juge + période + produit $+\varepsilon$

Construction: plan initial : $P = J + P \pmod 5$

Période

Juge					
	0	1	2	3	4
	1	2	3	4	0
	2	3	4	0	1
	3	4	0	1	2
	4	0	1	2	3

+ randomisation $\Rightarrow$

Période

Juge					
	0	1	2	4	3
	3	4	0	2	1
	2	3	4	1	0
	4	0	1	3	2
	1	2	3	0	4

Formalisation

Les étapes concrètes sont

- hétérogénéités identifiées à l'avance $\Rightarrow$ structure en blocs
- construction d'un plan initial "optimal" d
- randomisation par permutation aléatoire des unités $\Rightarrow$ plan final d_r

On note

- T ensemble des traitements $\mathbf{t}$
- U^* ensemble des unités virtuelles $\mathbf{u}$
- U ensemble des unités réelles $\mathbf{u}$
- Π groupe des permutations de U associé à la structure en blocs

Plan initial: $\left\{ \begin{array}{rcl} d & : & U^* \longrightarrow T \\ & & \mathbf{u} \longmapsto d(\mathbf{u}) \end{array} \right.$

Plan randomisé: $\left\{ \begin{array}{rcl} d_r & : & U \longrightarrow T \\ & & \mathbf{u} \longmapsto d_r(\mathbf{u}) = d\left(\pi^{-1}(\mathbf{u})\right) \end{array} \right.$

où π est une permutation aléatoire tirée dans Π

Au final, l'unité virtuelle $\mathbf{u}$ de U^*

- reçoit le traitement $d(\mathbf{u})$
- est affectée sur le terrain à l'unité réelle $\pi(\mathbf{u})$

Validité: cadre théorique

On postule un modèle de réponse **virtuelle** sur $\mathbf{u} \in U$:

$$y(\mathbf{t}, \mathbf{u}) = \tau_{\mathbf{t}} + \eta_{\mathbf{u}} \tag{1}$$

avec $\tau_{\mathbf{t}}$ effet fixe du traitement $\mathbf{t}$ et $\eta_{\mathbf{u}}$ effet de l'unité $\mathbf{u}$

$\Rightarrow$ hypothèses principales: 1) additivité; 2) $\eta_{\mathbf{u}}$ a des moments d'ordre ≤ 2 finis

La randomisation génère un modèle de réponse **observée** sur $\mathbf{u} \in U^*$:

$$Y_{\mathbf{u}} = \tau_{d(\mathbf{u})} + \eta_{\pi(\mathbf{u})} \tag{2}$$

qui peut s'écrire

$$Y_{\mathbf{u}} = \mu + \tau_{d(\mathbf{u})} + \varepsilon_{\mathbf{u}} \tag{3}$$

avec $\varepsilon_{\mathbf{u}}$ identiquement distribués, centrés et iso-corrélés

On a:

$$\begin{aligned} E(Y_{\mathbf{u}}) &= \tau_{d(\mathbf{u})} + \sum_{\mathbf{u}' \in U} P(\pi(\mathbf{u}) = u') \, . \, E(\eta_{\mathbf{u}'}) \\ Cov(Y_{\mathbf{u}}, Y_{\mathbf{v}}) &= \sum_{\mathbf{u}' \in U; \mathbf{v}' \in U} P(\pi(\mathbf{u}) = u'; \pi(\mathbf{v}) = v') \;\; E\left((\eta_{\mathbf{u}'} - \mu).(\eta_{\mathbf{v}'} - \mu)\right) \end{aligned}$$

On note

- $\mu = \frac{1}{N} \sum_{\mathbf{u}' \in U} E(\eta_{\mathbf{u}'})$
- $\sigma^2 = \frac{1}{N} \sum_{\mathbf{u}' \in U} E(\eta_{\mathbf{u}'} - \mu)^2$
- $\rho\sigma^2 = \frac{1}{N(N-1)} \sum_{\mathbf{u}', \mathbf{v}' \in U^2, \mathbf{u}' \neq \mathbf{v}'} E\left((\eta_{\mathbf{u}'} - \mu).(\eta_{\mathbf{v}'} - \mu)\right)$

Propriétés

- Si le groupe de permutations Π est transitif et si les tirages π dans Π sont équiprobables, alors $E(Y_{\mathbf{u}}) = \mu + \tau_{d(\mathbf{u})}$
- Si le groupe de permutations Π est doublement transitif et si les tirages π dans Π sont équiprobables, alors, pour $\mathbf{u} \neq \mathbf{u}'$
  - $Var(Y_{\mathbf{u}}) = \sigma^2$
  - $Cov(Y_{\mathbf{u}}, Y_{\mathbf{u}'}) = \rho \, \sigma^2$

Structures en blocs plus générales (distributives ou poset):

- facteurs blocs croisés et/ou hiérarchisés équilibrés
- Randomisation: enchaînement ordonné de permutations aléatoires et indépendantes entre les niveaux des différents facteurs blocs

Modèle *validé par la randomisation*

$$\begin{aligned} E(Y) &= X\tau, \\ Var(Y) &= \sum_k \xi_k S_k \end{aligned}$$

où les ξ_k sont des paramètres à estimer et les S_k

- sont des matrices de projection sur des sous-espaces mutuellement orthogonaux appelés *strates*
- se calculent à partir de la structure en blocs.

$\Longrightarrow$ analyse par le modèle linéaire mixte (ML, REML) ou par moindres carrés en projetant dans certaines strates

À retenir:

À retenir:

- La randomisation est ici conçue comme une permutation aléatoire des unités expérimentales, par tirage dans un groupe de permutations associé à la structure en blocs souhaitée.
- Sous des hypothèses très faibles sur $y(\mathbf{t}, \mathbf{u})$ (additivité, moments finis d'ordre 2), le plan complètement randomisé permet de valider jusqu'au 2nd ordre le modèle linéaire d'analyse de la variance.
- Ce résultat se généralise à des structures en blocs très générales et conduit à un modèle linéaire mixte dont la structure de la matrice de variance-covariance est connue.
- Pour des petits plans, il existe des méthodes pour éviter des répartitions non désirables (randomisation restreinte).

Plan

Essai contrôlé/comparatif randomisé (ECR; RCT)

"Gold standard" de la médecine fondée sur les preuves

- s traitements à comparer
- recrutement de N sujets
- répartition aléatoire des participants entre les "groupes" correspondant à chaque traitement

Objectifs de la randomisation

- minimiser les biais de confusion d'effets avec des variables non contrôlées
- minimiser les biais de sélection et faciliter les démarches en aveugle (non prédictibilité)
- contribuer à la démarche inférentielle

Procédures possibles dans le cas séquentiel ($s = 2$)

- randomisation simple, ou complète: probabilités fixées pour tous les sujets
  totalement non prédictible mais risque de déséquilibre entre traitements si N petit
- randomisation restreinte: probabilités adaptatives favorisant le traitement minoritaire

Exemple: plan randomisé à dé pipé (accelerated biased coin design; Baldi Antognoni et Giovagnoli, 2004)

$$\begin{aligned} P_j &= 1/2 && \text{si} \quad D_{j-1} = 0, \\ &= \frac{\left|D_j - 1\right|^a}{\left|D_j - 1\right|^a + 1} && \text{si} \quad D_{j-1} \leq -1, \\ &= \frac{1}{\left|D_j - 1\right|^a + 1} && \text{si} \quad D_{j-1} \geq -1, \end{aligned}$$

avec $D_{j-1} = N_A(j-1) - N_B(j-1)$

À retenir:

Pour des plans de petite taille,

- le choix de la procédure de randomisation a des conséquences notables sur
  - la prédictibilité des séquences (donc le risque de biais de sélection),
  - les effectifs des groupes affiliés aux traitements (donc l'efficacité du plan),
- il joue également sur les taux effectifs des erreurs de 1e et 2e espèce
- il existe des comparaisons détaillées.

cf. Rosenberger, 2015 (s'intéresse au cas $N \simeq 50$)

Plan

Rappel sur les tests de randomisation, par l'exemple

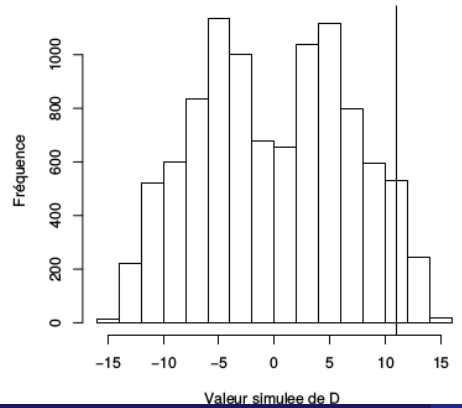
Supposons par exemple que l'on a observé, de façon indépendante, les huit rendements suivants sur chacune des variétés A et B :

```
A : 51.58 63.88 59.56 61.88 58.34 52.95 50.11 58.40

B : 72.05 73.21 84.53 76.44 67.28 78.78 71.40 63.20
```

La différence observée entre les médianes[20] des deux variétés est égale à 14,26 q/ha. En effectuant 10000 permutations aléatoires des 16 données et en réattribuant les 8 premières valeurs à A et les 8 dernières à B, on obtient la distribution de la Figure 9 pour la différence entre médianes, avec une valeur limite égale à 11 q/ha pour un test unilatéral avec un risque de première espèce de 5%[21]. La valeur observée dépasse la valeur limite, on rejette donc l'hypothèse H_0 d'absence de différence entre les variétés.

FIG. 9 – Test de permutation : distribution de la différence des médianes entre les variétés A et B, pour 10000 permutations aléatoires des données. Le trait vertical correspond à la valeur limite pour un risque de 1ère espèce de 5%



À retenir

- la randomisation permet de justifier (partiellement) l'analyse des résultats par l'anova et le modèle linéaire
  - postulat essentiel: additivité entre effets traitements et effets unités $y(\mathbf{t},\mathbf{u}) = \tau_{\mathbf{t}} + \eta_{\mathbf{u}}$
- elle permet aussi d'appliquer des méthodes d'inférence "model-free"
  - pas de postulat d'additivité: $y(\mathbf{t},\mathbf{u}) - y(\mathbf{t}',\mathbf{u})$ peut dépendre de u
  - hypothèses à tester (typiquement)
    - H_0 dite de Fisher: $\forall \mathbf{u}, \mathbf{t} \neq \mathbf{t}' : y(\mathbf{t},\mathbf{u}) = y(\mathbf{t}',\mathbf{u})$
    - H_0 dite de Neyman: $\forall \mathbf{t} \neq \mathbf{t}' : y(\mathbf{t},\bullet) = y(\mathbf{t}',\bullet)$
  - sans plus d'hypothèses sur $y(\mathbf{t},\mathbf{u})$, H_0 peut être testée par un **test de randomisation**
- ces méthodes sont intéressantes pour des petits plans, puisque les raisonnements asymptotiques ne tiennent pas

En guise de conclusion . . . historique

En 1935 la présentation par Jerzy Neyman, à la *Royal Statistical Society*, de l'article *Statistical Problems in Agricultural Experimentation* est à l'origine d'une controverse de longue haleine entre Jerzy Neyman et Ronald A. Fisher.

Cette controverse est en partie liée aux relations entre randomisation, modélisation et inférence sur la causalité.

Certaines de ces questions font l'objet des travaux méthodologiques les plus récents sur la randomisation (cf. Ping 2017 et discussion).

Plan

Références

- Bailey R.A. (2008). *Design of Comparative Experiments*. Cambridge University Press, Cambridge.
- Bailey R.A. (1991). Strata for randomized experiments (with discussion). *Journal of the Royal Statistical Society Series B*, **53**, 27–78.
- Bardin A. et Azaïs J.-M. (1990). Une hypothèse minimale pour une théorie des plans d'expériences randomisés. *Revue de Statistique Appliquée*, **38**, 5–20.
- Ding P., Dasgupta T. (2017). A paradox from randomization-based causal inference (with discussion). *Statistical Science*, **32**, 331-345.
- Fisher, R. A. (1935, 1971). *The Design of Experiments*. (1st Ed. Oliver and Boyd, Edinburg, 1935; 9th ed. Hafner Publishing Company New York, 1971).

- Neyman J. (1935). Statistical problems in agricultural experimentation (with discussion). *Suppl. J. Roy. Statist. Soc. Ser. B* **2**, 107-180.
- Rosenberger W.F. (2015). Randomization in small clinical trials. In: *Randomization for small clinical trials* seminar (6-10 July 2015) of the Isaac Newton Institute for Mathematical Sciences, Cambridge.
- Sabbaghi A., Rubin D.B. (2014). Comments on the Neyman-Fisher controversy and its consequences. *Statistical Science*, **29**, 267-284.
- `https://en.wikipedia.org/wiki/Randomized_experiment`
- `https://fr.wikipedia.org/wiki/Essai_randomisé_contrôlé`

Packages R

Packages R pour construire des plans factoriels réguliers: `FrF2, planor (INRA)`

Caractéristiques de `planor`:

- construit et randomise des plans factoriels réguliers
- en particulier des plans factoriels **fractionnaires**
- pour expériences réelles ou numériques
- fonction de randomisation pour des structures en blocs généralisées

```
> library(planor)
> Design <- data.frame(block=rep(1:4,rep(2,4)),
+         treatment=c("A1","B1","A2","B2","A3","B3","A4","B4"))
> Design
  block treatment
1     1        A1
2     1        B1
3     2        A2
4     2        B2
5     3        A3
6     3        B3
7     4        A4
8     4        B4
> planor.randomize(~block/UNITS, data=Design)
  block treatment
1     1        A2
2     1        B2
3     2        B4
4     2        A4
5     3        B3
6     3        A3
7     4        A1
8     4        B1
```

```
> RowColDes <- data.frame(row=rep(1:3,rep(3,3)),col=rep(1:3,3),treatment=LETTERS[c(1:3,2,3,1,3,1,2)])
> RowColDes
  row col treatment
1   1   1         A
2   1   2         B
3   1   3         C
4   2   1         B
5   2   2         C
6   2   3         A
7   3   1         C
8   3   2         A
9   3   3         B
> planor.randomize(~row*col, data=RowColDes)
  row col treatment
1   1   1         B
2   1   2         A
3   1   3         C
4   2   1         C
5   2   2         B
6   2   3         A
7   3   1         A
8   3   2         C
9   3   3         B
```