

Plans d'expérience

Applications en recherche pré-clinique

Christophe Lalanne

ch.lalanne@mac.com

www.aliquote.org

Plan de l'exposé

rappels sur les principaux types de plans d'expérience • principe de l'analyse de variance • procédures de comparaisons multiples

TL;DR

Lorsque les conditions d'application sont vérifiées, le test de Student reste le plus puissant pour comparer deux moyennes. Les plans d'expérience permettent d'arranger les sources de variation de manière à maximiser la précision de l'estimation d'un effet, c'est le principe de l'analyse de variance.

Plans d'expérience

To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination : he may be able to say what the experiment died of.

— Ronald Fisher

The Environment and Disease : Association or Causation ?

Criterion	Description	Limitations
1. Strength	A strong effect size for an association	This reduces the chance of minor unmeasured confounding, but assumes that major confounding factors have been accounted for. Weak associations may also be causal.
2. Consistency	Repeated observations of an association in different populations/circumstances	This assumes that all necessary causal factors are evenly distributed between populations. If a risk factor-outcome association were present only in men, would this imply non-causality?
3. Specificity	A risk factor leads to a single outcome	There is no reason why a risk factor should be associated with a single disorder (e.g. multiple disorders associated with alcohol misuse).
4. Temporality	The cause should precede the effect	A study should ideally demonstrate this. However the fact that one event follows another does not rule out the opposite direction of causation on other circumstances. For example, depression may cause physical ill-health but the opposite may also occur.
5. Biological gradient	A 'dose-response' relationship	This assumes that the 'ceiling' of risk has not been reached. A single life-event may be sufficient to cause depression with no influence of further events. Of little use for cross-sectional associations since a 'dose-response' pattern of association would be predicted with either direction of causation.
6. Plausibility	That the hypothesis is biologically plausible	Frequently a highly subjective judgement, given the volume of the biological literature. There are many historical examples of important findings rejected on the ground of implausibility at the time (e.g. Darwin's theory of the Origin of Species).
7. Coherence	That the interpretation does not conflict with the known biology of the disease	This depends heavily on the quality of the ancillary information. It also is not entirely consistent with the principle of hypothesis refutation.
8. Experimental evidence	Evidence from interventional research	Interventional research may not be ethical and/or feasible for many cause effect investigations. The intervention may not be discrete enough to infer causation.
9. Analogy	Similar associations in other fields	A highly subjective judgement.

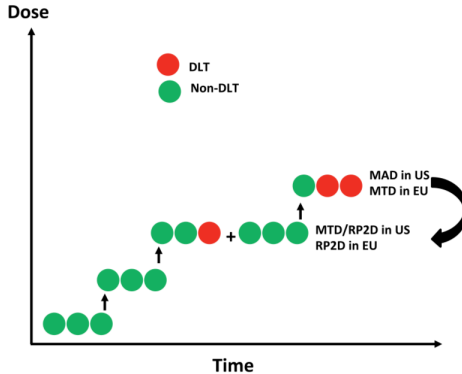
Principe des plan d'expérience

Mise en œuvre organisée d'un ensemble d'unités expérimentales pour caractériser l'effet de certains traitements, ou combinaison de traitements, sur une ou plusieurs variables réponses.

Prise en considération d'un ou plusieurs facteurs de nuisance pendant la constitution du dessin expérimental : organiser les sources de variation indésirables de façon à ce qu'elles affectent les traitements de manière équivalente, rendant ainsi possible la comparaison entre traitements.

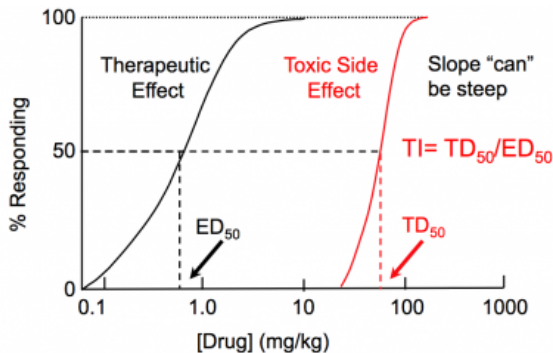
- Minimiser le nombre d'essais et maximiser la précision
- Respecter des contraintes externes (budget, sujets, éthique, durée, exposition, environnement, etc.)

Etudes pré-cliniques : cas d'un plan 3 + 3



Protocole d'escalade de dose¹

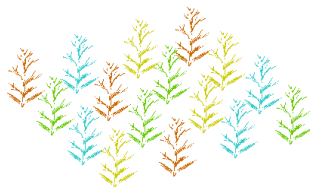
Etudes pré-cliniques : plan dose-réponse



Protocole de recherche d'effet thérapeutique²

Plan complet randomisé³ (CRD)

- Les répliques de traitement sont administrées aléatoirement et de manière indépendante aux sujets
- Les sujets adjacents peuvent recevoir le même traitement
- Le cas à deux traitements correspond au test de Student pour deux échantillons indépendants



4 traitements (A-D), 4 répliques (1-4)

A1 B1 C1 A2

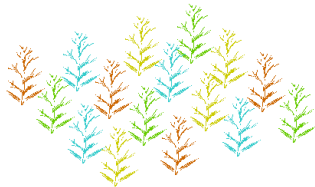
D1 A3 D2 C2

B2 D3 C3 B3

C4 A4 B4 D4

Blocs complets aléatoires (RCB)

- Traitements assignés au hasard à l'intérieur des blocs, une seule fois par bloc
- Le nombre de blocs est égal au nombre de répliques
- N'importe quel traitement peut être adjacent à un autre entre blocs mais pas à l'intérieur des blocs
- Permet de contrôler la **variabilité due aux effets spatiaux**



4 traitements (A-D), 4 blocs (I-IV)

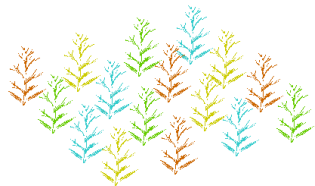
Block I A B C D

Block II D A B C

Block III B D C A

Block IV C A B D

- Traitements assignés aléatoirement et de manière unique par ligne et par colonne
- Même nombre de lignes, colonnes et traitements
- Intéressant pour **contrôler la variabilité dans deux directions** (souvent orthogonales)



4 traitements (A–D), 4 lignes (I–IV) et 4 colonnes (1–4)

Row I A B C D

Row II C D A B

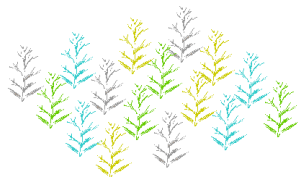
Row III D C B A

Row IV B A D C

Column 1 2 3 4

Plan factoriel 2^2 CRD

- Les traitements correspondent au croisement des deux niveaux de deux facteurs dans un plan complet
- Les répliques sont assignées au hasard aux sujets indépendants
- Les sujets adjacents peuvent recevoir la même combinaison de traitement



4 combinaisons de traitement (Aa, Ba, Ab et Bb) et 4 répliques (1–4)

Aa1 Ba1 Ab1 Aa2

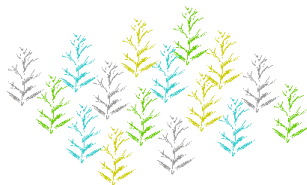
Bb1 Aa3 Bb2 Ab2

Ba2 Bb3 Ab3 Ba3

Ab4 Aa4 Ba4 Bb4

Plan factoriel 2^2 RCB

- Les traitements correspondent au croisement des deux niveaux de deux facteurs dans un plan de blocs aléatoires
- Les traitements sont assignés au hasard et de manière unique à l'intérieur des blocs de sujets adjacents
- Le nombre de blocs est égal au nombre de répliques
- N'importe quel traitement peut être adjacent à un autre entre blocs mais pas à l'intérieur des blocs



4 blocs et 4 combinaisons de traitement
(Aa, Ba, Ab et Bb)

Block IV Aa Ba Ab Bb

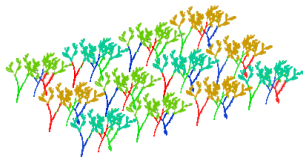
Block III Bb Aa Ba Ab

Block II Ba Bb Ab Aa

Block I Ab Aa Ba Bb

Plan split-plot CRD

- Principe du CRD dans lequel on introduit un niveau additionnel de stratification pour des sujets indépendants (“subplot”) auxquels on administre aléatoirement un autre type de traitement
- Les répliques sont assignées au hasard aux sujets expérimentaux du premier niveau, pas au subplot
- Les sujets indépendants adjacents peuvent recevoir le même traitement

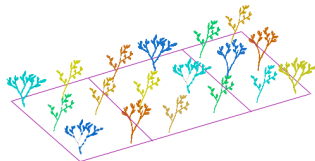


4 répliques de 3 traitements (A–C)
divisés en 3 sous-traitements (a–c)

A1a A1b A1c B1c B1b B1a A2b A2c A2a
C1a C1c C1b
C2c C2a C2b A3b A3c A3a B2c B2a B2b
C3b C3a C3c
B3b B3c B3a A4a A4c A4b C4c C4a C4b
B4a B4b B4c

Plan split-block RCB

- Deux ensembles de traitement (possiblement en gradient) sont randomisés entre eux à l'intérieur de bandes dans un plan en blocs aléatoires
- Utilisable lorsque les traitements doivent être administrés intégralement entre chaque bloc
- Le nombre de blocs est égal au nombre de répliques



3 blocs (I-III), 3 + 2 traitements (A-C, 1-2)

Block I Block II Block III

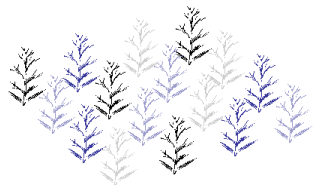
A1 A2 C2 C1 B1 B2

B1 B2 A2 A1 C1 C2

C1 C2 B2 B1 A1 A2

Régression linéaire simple

- Les traitements sont déterminés et fixés à partir de quantités numériques choisies dans un ensemble de valeurs admissibles
- Les répliques sont assignées complètement au hasard aux sujets indépendants
- Les sujets adjacents peuvent recevoir le même traitement



4 traitements (0, 1, 2, 5) et 4 répliques (a-d)

5a 2a 0a 5b

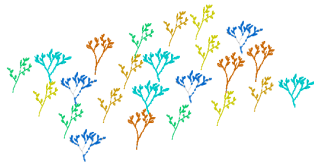
1a 5c 1b 0b

2b 1c 0c 2c

0d 5d 2d 1d

Régression linéaire avec groupes indépendants

- Les traitements sont déterminés et fixés à partir de quantités numériques choisies dans un ensemble de valeurs admissibles, et croisés avec les niveaux d'un facteur
- Les répliques sont assignées complètement au hasard aux sujets indépendants



3 niveaux (1, 2, 5), 2 traitements (A, B) et
4 répliques (a–d)

5Aa 2Aa 5Ba 5Ab 2Ba 1Ba

1Bb 1Aa 2Bb 2Ab 1Bc 1Ab

5Ac 2Bc 2Ac 1Ac 5Bb 5Bc

1Ad 5Bd 5Ad 1Bd 2Bd 2Ad

Blocs complets

Pour des raisons d'économie de place, les animaux utilisés pour une expérience sont mis dans des cages qui peuvent en contenir chacune 12. On dispose au total de 10 cages et l'on doit comparer 3 traitements A, B et C. L'une des solutions consiste à répartir au hasard les 3 traitements parmi les 120 animaux. On conçoit cependant qu'il puisse être intéressant, si l'on pense que les résultats observés dans chaque case risquent de présenter une certaine homogénéité par rapport à l'ensemble, et qu'il existe corrélativement une certaine hétérogénéité d'une cage à l'autre, d'attribuer les 3 traitements à 4 animaux à l'intérieur de chaque cage (par tirage au sort).

Blocs incomplets équilibrés

Comparaison des réactions cutanées locales de 5 traitements A, B, C, D, E chez la souris. On souhaiterait utiliser l'animal comme son propre témoin, en d'autres termes prendre chaque souris comme bloc. Il se peut cependant que, compte-tenu des dimensions de la souris, il ne soit possible d'appliquer que 4 traitements à chaque animal.

Carré latin

Comparaison de 4 traitements appliqués localement au même animal, mais à des endroits différents. Si le lieu d'injection est une cause systématique et importante de variation, on constituera des blocs "animal-lieu d'injection" avec une seule unité expérimentale (Lellouch & Lazar, 1974).

```
1 library(agricolae)
2 tx <- LETTERS[1:4]
3 p1 <- design.crd(trt = tx, r = rep(4, 4), seed = 101)    ## CRD
4 p1$book
5 p2 <- design.rcbd(trt = tx, r = 4, seed = 121)          ## RCB
6 tx1 <- paste0("c", 1:4)
7 tx2 <- paste0("r", 1:4)
8 p3 <- design.graeco(tx1, tx2, serie = 1)                ## Carré latin
```

Modèles statistiques courants

Cas de la comparaison de deux moyennes

Test de Student

Pour deux échantillons indépendants, la statistique de test se définit ainsi :

$$t_{\text{obs}} = \frac{\bar{x}_1 - \bar{x}_2}{s_c \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad s_c = \left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \right)^{1/2},$$

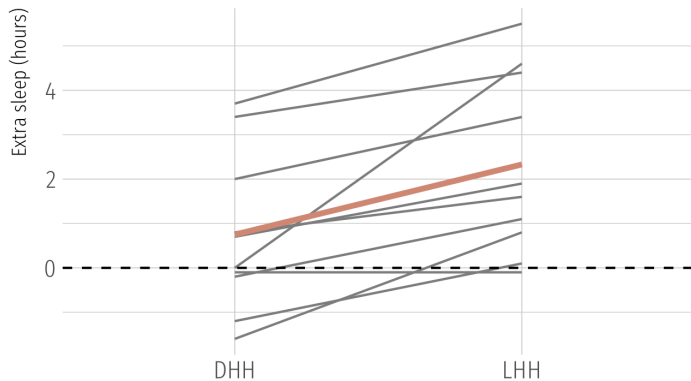
où les \bar{x}_i et n_i sont les moyennes et effectifs des deux échantillons, et s_c est la variance commune pour la différence de moyennes d'intérêt. Sous H_0 , cette statistique de test suit une loi de Student à $n_1 + n_2 - 2$ degrés de liberté.

Un intervalle de confiance à $100(1 - \alpha)\%$ pour la différence $\bar{x}_1 - \bar{x}_2$ peut être construit comme suit :

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha, n_1 + n_2 - 2} s_c \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

avec $P(t < t_{\alpha, n_1 + n_2 - 2}) = 1 - \alpha/2$.

Illustration



Données sur les hypnotiques (Gosset, 1908)

```
1 t.test(extra ~ group, data = sleep)
2 t.test(extra ~ group, data = sleep, paired = TRUE)
```

Résultat du test de Student pour échantillons appariés

Paired t-test

data: extra by group

t = -4.0621, df = 9, p-value = 0.002833

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

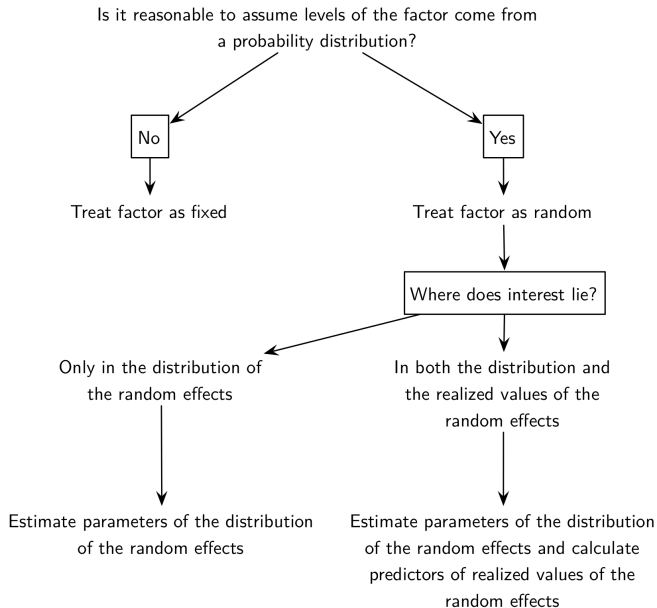
-2.4598858 -0.7001142

sample estimates:

mean of the differences

-1.58

Facteurs à effets fixes ou aléatoires



L'ANOVA constitue une extension naturelle au cas où plus de deux moyennes de groupe sont à comparer. Pour k échantillons, l'hypothèse nulle se lit :

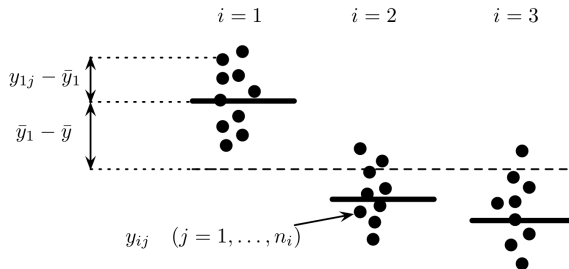
$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k,$$

et l'hypothèse alternative (H_1) est l'existence d'au moins une paire de moyennes qui diffèrent (négation logique de H_0).

Si l'on exprime chaque observation comme une déviation par rapport à sa propre moyenne de groupe, $y_{ij} = \bar{y}_i + \varepsilon_{ij}$, on voit que la variabilité totale peut se décomposer comme suit :

$$\underbrace{(y_{ij} - \bar{y})}_{\text{totale}} = \underbrace{(\bar{y}_i - \bar{y})}_{\text{groupe}} + \underbrace{(y_{ij} - \bar{y}_i)}_{\text{résiduelle}}.$$

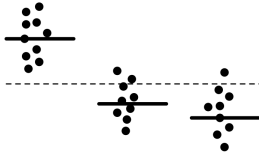
Illustration



Exemple de distribution sur 3 groupes

Conditions de validité

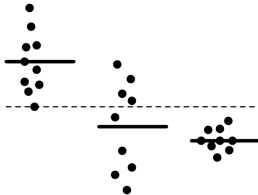
Variances homogènes



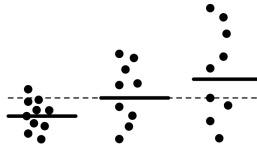
Moyennes constantes



Variances non-homogènes



Variance dépendant de la moyenne



Indépendance, égalité des variances, normalité des résidus

Le modèle d'ANOVA

Soit y_{ij} la j^{e} observation dans le groupe i (facteur A). On peut résumer le modèle à effet comme suit :

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

où μ désigne la moyenne générale, α_i l'effet du groupe i ($i = 1, \dots, a$), et $\varepsilon_{ij} \sim N(0, \sigma^2)$ un terme d'erreur aléatoire. On impose généralement que $\sum_{i=1}^a \alpha_i = 0$.

L'hypothèse nulle se lit $H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_a$, et se teste à l'aide d'un test F à $a - 1$ et $N - a$ degrés de liberté.

Des méthodes graphiques (boxplot ou QQ-plot) sont recommandées pour vérifier les hypothèses de normalité ou d'homogénéité des variances à la place de tests formels (Levene, Fisher, Bartlett, etc.).

Il existe des alternatives au modèle paramétrique de base : ANOVA de Welch ne supposant pas l'égalité des variances (Welch, 1951), ANOVA non paramétrique de Kruskal-Wallis, ANOVA robuste avec utilisation de moyennes tronquées (Wilcox & Keselman, 2003).

Le modèle d'ANOVA

Soit y_{ijk} la k^{e} observation pour le niveau i du facteur A ($i = 1, \dots, a$) et le niveau j du facteur B ($j = 1, \dots, b$). Le modèle complet avec interaction s'écrit

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},$$

où μ désigne la moyenne générale, α_i (β_j) l'écart à la moyenne des moyennes de groupe pour le facteur A (B), γ_{ij} les écarts à la moyenne des moyennes pour les traitements $A \times B$, et $\varepsilon_{ijk} \sim N(0, \sigma^2)$ la résiduelle. Les effets α_i et β_j sont appelés effets principaux, tandis que γ_{ij} est l'effet d'interaction.

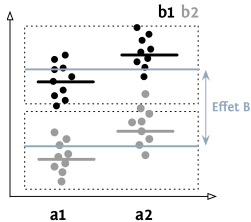
Les hypothèses nulles associées au modèle complet sont :

1. $H_0^A : \alpha_1 = \alpha_2 = \dots = \alpha_a, (a - 1) \text{ ddl}$
2. $H_0^B : \beta_1 = \beta_2 = \dots = \beta_b, (b - 1) \text{ ddl}$
3. $H_0^{AB} : \gamma_{11} = \gamma_{13} = \dots = \gamma_{ab}, (a - 1)(b - 1) \text{ ddl}$

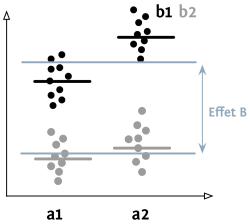
Des tests F (CM effets / CM résiduelle) permettent de tester ces hypothèses.

Interpréter une interaction

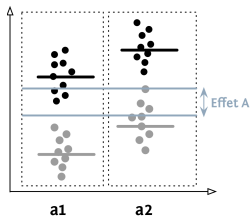
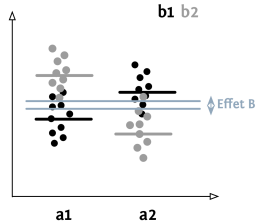
Absence d'interaction



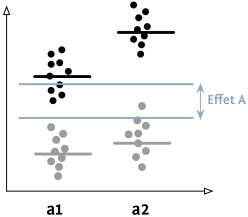
Interaction ordonnée



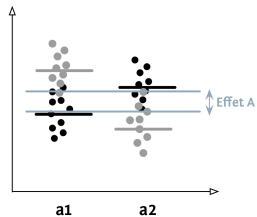
Interaction croisée



L'effet de B est le même quel que soit le niveau de A.



L'effet de B est plus important dans la condition a2.



L'effet de B s'inverse entre a1 et a2.

Remarque sur les formules R

R suit les conventions de notation proposées par Wilkinson & Rogers pour les plans d'expérience (Chambers & Hastie, 1992 ; Wilkinson & Rogers, 1973).

y ~

x régression linéaire simple

x + 0 idem avec suppression de l'intercept

a + b deux effets principaux (croisement)

a * b équivalent à 1 + a + b + a :b (interaction)

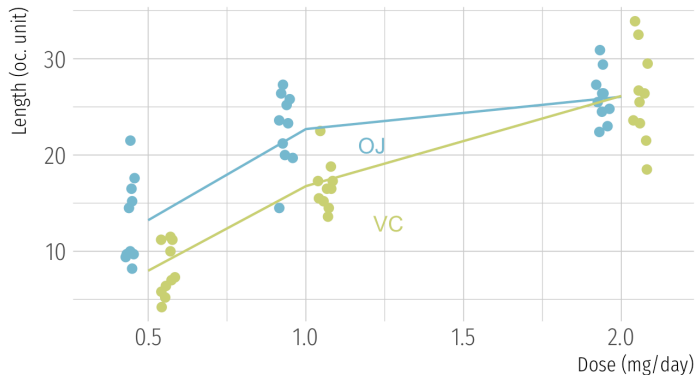
a / b équivalent à 1 + a + b + a %in% b (emboîtement)

```
fm <- y ~ a * b * c            ## (A, B, C, AB, AC, BC, ABC)
m1 <- aov(fm, data = d)        ## estimation
update(mod1, . ~ . -a:b:c)    ## remove interaction ABC
```

Illustration

The response is the length of odontoblasts in 10 guinea pigs at three dose levels of Vitamin C with two delivery methods (orange juice or ascorbic acid).

— Bliss (1952)




```

1 ToothGrowth$dose <- factor(ToothGrowth$dose)
2 fm <- lm ~ supp * dose
3 replications(fm, data = ToothGrowth)
4 aggregate(fm, ToothGrowth, mean)
5 m <- aov(fm, data=ToothGrowth)
6 summary(m)
7 model.tables(m, type = "means", se = TRUE, cterms = "supp:dose")

```

Tableau d'ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
supp	1	205.4	205.4	15.572	0.000231	***
dose	2	2426.4	1213.2	92.000	< 2e-16	***
supp:dose	2	108.3	54.2	4.107	0.021860	*
Residuals	54	712.1	13.2			

1. Les tests précédents n'indiquent pas quelles paires de moyennes diffèrent significativement, mais permettent de se prononcer sur l'existence d'un effet et le rejet de l'hypothèse nulle associée. Pour préciser quels sont les traitements qui diffèrent deux à deux, il faudrait utiliser des procédures (post-hoc) de comparaisons multiples.
2. Le facteur dose est traité comme une variable qualitative non ordonnée. Un test de linéarité de la relation $\text{len} \sim \text{dose}$ serait toutefois intéressant. Sous R, on peut réaliser un tel test par régression ou en utilisant des contrastes polynomiaux.
3. Pour tester formellement l'hypothèse d'égalité des variances, il faut travailler au niveau de tous les traitements :

```
bartlett.test(len ~ interaction(supp, dose), data = ToothGrowth)
```

Comparaisons multiples planifiées ou non

1. Méthode des contrastes

Avec k échantillons, on peut définir $k - 1$ contrastes orthogonaux

$$\phi = \sum_{i=1}^k c_i \bar{x}_i, \quad \sum_i c_i = 0 \text{ et } \phi_u^t \phi_v = 0$$

et utiliser comme statistique de test, $\frac{\phi}{s_\phi}$, où $s_\phi^2 = s^2 \sum_i \frac{c_i^2}{n_i} \sim t(\nu)$.

2. Tests post-hoc

On a un ensemble de $k(k - 1)/2$ paires de moyennes. Pour $k = 4$, en fixant $\alpha = 0.05$, le risque d'ensemble ou FWER (en supposant les tests indépendants) devient $1 - (1 - 0.05)^6 = 0.265$, soit 27 % de chance de déclarer à tort un résultat significatif.

Deux stratégies, généralement conservatrices : modifier la statistique de test (Tukey HSD) ou modifier le risque de première espèce (Bonferroni) (Christensen, 2002). Autres approches : méthode adaptative (Holm), contrôle du FDR, permutations, etc. (Dudoit & Laan, 2008).

Lorsque les effectifs ne sont pas tous égaux entre les traitements, les choses se compliquent, notamment au niveau du calcul séquentiel des sommes de carrés (Herr, 1986).

Cas d'un plan à deux facteurs, A et B :

- Type I (défaut) : $SS(A)$, $SS(B|A)$, puis $SS(AB|B, A)$
- Type II : $SS(A|B)$, puis $SS(B|A)$ (pas d'interaction)
- Type III : $SS(A|B, AB)$, $SS(B|A, AB)$ (interpréter chaque effet principal après avoir pris en compte les autres effets principaux et l'interaction)

Méthode d'ajustement *a posteriori*

Applicable également dans le cas des plans de type avant/après, l'analyse de covariance fournit le meilleur estimateur pour la différence différence entre deux groupes de sujets, en comparaison de la simple analyse des scores de différence ou d'une approche par modèle mixte (Senn, 2006).

Modèle d'ANCOVA

Soit y_{ij} la j ème observation dans le groupe i . À l'image du modèle d'ANOVA à un facteur, le modèle d'ANCOVA s'écrit :

$$y_{ij} = \mu + \alpha_i + \beta(x_{ij} - \bar{x}) + \varepsilon_{ij},$$

où β est le coefficient de régression liant la réponse y et le cofacteur x (continu), avec \bar{x} la moyenne générale des x_{ij} , et toujours un terme d'erreur $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

Notons que l'on fait l'hypothèse que β est le même dans chaque groupe. Cette hypothèse de parallélisme peut se vérifier en testant la significativité du terme d'interaction $\alpha\beta$.

La réponse moyenne ajustée pour l'effet du co-facteur numérique s'obtient simplement comme $\bar{\alpha}_i + \hat{\beta}(\bar{x}_i - \bar{x})$, où \bar{x}_i est la moyenne des x dans le i ème groupe.

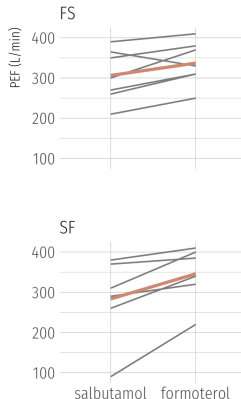
Étude de cas : essai croisé

Utilisation d'un modèle à effet aléatoire

Two-period AB-BA randomised crossover trial in which patients are treated with two bronchodilators, salbutamol (S) and formoterol (F). The outcome is the peak expiratory flow (PEF). Patients received F-S or S-F.

— Senn (1993)

ID	seq	S	F
01	FS	270	310
02	SF	370	385
03	SF	310	400
04	FS	260	310
05	SF	380	410
06	FS	300	370
07	FS	390	410
09	SF	290	320
10	FS	210	250
...			



```
1 library(nlme)
2 d <- read.csv("peakflow.csv")
3 m <- lme(value ~ treatment + period + sequence, data = d,
4         random = ~ 1 | patient)
5 summary(m)
```

Utilisation de sommes de carré de type III :

```
1 library(lme4)
2 library(lmerTest)
3 m <- lmer(value ~ treatment + period + sequence + (1 | patient),
4         data = d)
5 anova(m)
```


Modèle d'ANOVA

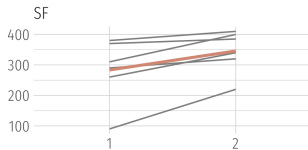
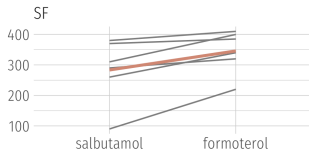
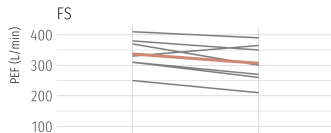
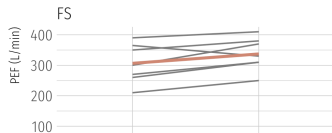
Analysis of Variance Table of type III with Satterthwaite approximation for degrees of freedom

	Sum Sq	Mean Sq	NumDF	DenDF	F.value	Pr(>F)
treatment	14035.9	14035.9	1	11	18.7044	0.001205 **
period	1632.1	1632.1	1	11	2.1749	0.168314
sequence	24.1	24.1	1	11	0.0321	0.861076

On obtiendrait les mêmes résultats pour l'effet séquence en utilisant la formule suivante :

```
1 m <- lmer(value ~ treatment * period + (1 | patient), data = d)
2 anova(m)
```

Effet séquence = effet d'interaction

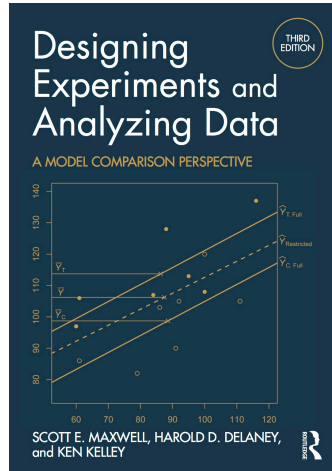
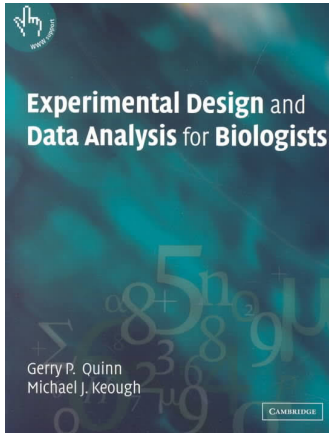


Termes ou effets à considérer : traitement (critère de jugement, $\alpha_S - \alpha_F$), période ($\pi_2 - \pi_1$), carry-over ($\lambda_F - \lambda_S$ —en principe, on peut supposer qu'il n'y en a pas car les mesures sont réalisées 8h après traitement), individuel, résiduelle.

	Période 1	Période 2
FS	$\mu + \pi_1 + \alpha_F$	$\mu + \pi_2 + \alpha_S + \lambda_F$
SF	$\mu + \pi_1 + \alpha_S$	$\mu + \pi_2 + \alpha_F + \lambda_S$

Les facteurs sujet et période sont considérés comme des blocs et on a un RCB dans chacun des deux cas. Pour l'effet séquence (SF ou FS), il s'agit d'un plan en blocs incomplets. Le problème avec le dessin AB/BA est que l'**effet traitement est contaminé par l'effet carry-over** ($\mathbb{E}(\alpha_S - \alpha_F) = \alpha_S - \alpha_F + 0.5(\lambda_F - \lambda_S)$).

Pour aller plus loin



Références I

- Bliss, C. I. (1952). *The Statistics of Bioassay : With Special Reference to the Vitamins*. (S.I.) : Academic Press.
- Chambers, J. M., & Hastie, T. J. (Éds). (1992). *Statistical Models in S*. (S.I.) : Wadsworth & Brooks.
- Christensen, R. (2002). *Plane Answers to Complex Questions : The Theory of Linear Models*. New York : Springer.
- Dudoit, S., & Laan, J. van der. (2008). *Multiple Testing Procedures with Applications to Genomics*. New York : Springer.
- Gosset, W. S. (1908). The Probable Error of a Mean. *Biometrika*, 6(1), 1–25.
- Herr, D. (1986). On the History of Anova in Unbalanced, Factorial Designs : The First 30 Years. *The American Statistician*, 40(4), 265–270.
- Lellouch, J., & Lazar, P. (1974). *Méthodes statistiques en expérimentation biologique*. Paris : Flammarion.
- Senn, S. (2006). Change From Baseline and Analysis of Covariance Revisited. *Statistics in Medicine*, 25(24), 4334–4344.
- Senn, S. S. (1993). *Cross-over Trials in Clinical Research*. Chichester, England : John Wiley & Sons.
- Welch, B. L. (1951). On the Comparison of Several Mean Values : an Alternative Approach. *Biometrika*, 38, 330–336.

- Wilcox, R. R., & Keselman, H. J. (2003). Modern Robust Data Analysis Methods : Measures of Central Tendency. *Psychological Methods*, 8(3), 254–274.
- Wilkinson, G. N., & Rogers, C. E. (1973). Symbolic Description of Factorial Models for Analysis of Variance. *Applied Statistics*, 22, 392–399.