CH 04. Puissance et réplication

Contexte

Comment déterminer un nombre approprié de répétitions pour une expérience à réaliser ? Le nombre d'observations recherché est celui qui nous permet de mettre en évidence de manière significative une différence entre deux niveaux du facteur d'intérêt.

Ce que l'on sait:

Plus de répétitions signifie plus d'information et une capacité à déclarer significatives des différences entre traitements plus petites
Moins de répétitions et/ou une variabilité résiduelle importante signifie que l'on risque de passer à côté de différences entre traitements.

Plusieurs éléments interviennent dans la décision:

Les ressources dont on dispose
Le choix des traitements [c'est un autre aspect de la planification]
Risques associés avec des décisions incorrectes
L'amplitude des différences entre traitements que l'on veut détecter comme statistiquement significative
Variabilité associée avec les mesures et les unités expérimentales

Plan

1. Rappels d'inférence sur une différence entre deux moyennes
2. Choix du nombre de réplications sur base de la plus petite différence significative
3. Rappels sur la notion de puissance d'un test
4. Choix du nombre de réplications basé sur la puissance
5. Différents niveaux de réplication

Erreur standard sur la différence

Variance d'une différence entre deux niveaux d'un facteur (fixe) :

$$var(\mu_2 - \mu_1) = var(\mu_2) + var(\mu_1) = 2var(\mu_i)$$

Elle est toujours supérieure à la variance d'une moyenne ($var(\mu_i)$) car la différence combine les erreurs (indépendantes) d'estimation des deux moyennes

Dans le cas d'un modèle fixe (Biométrie, Ch. 3 ANOVA1)

$$var(\mu_i) = \sigma^2/n$$

On l'estime avec S^2 ou MSE

Dès lors:

$$var(\hat{\mu}_2 - \hat{\mu}_1) = var(\hat{\mu}_2) + var(\hat{\mu}_1) = 2\,S_Y^2/n$$

On définit ainsi l'erreur standard sur la différence entre deux traitements:

$$SED = \sqrt{2\,S_Y^2/n}$$

Plus petite différence significative

Pour tester l'égalité des moyennes μ_1 et μ_2 de deux traitements, on utilise le rapport suivant:

$$t = (\hat{\mu}_2 - \hat{\mu}_1)/SED$$

qui suit une distribution de Student avec N-m degrés de libertés (ce sont les d.l. associés à MSE, l'estimateur de σ^2 et, indirectement, de SED).

On rejette l'hypothèse d'égalité des moyennes (au niveau α) lorsque

$$|t| > t_{N-m\,;\,1-\alpha/2}$$

On peut ainsi définir la plus petite différence significative :

$$LSD = t_{N-m\,;\,1-\alpha/2} SED = t_{N-m\,;\,1-\alpha/2} \sqrt{2\,S_Y^2/n}$$

telle que toute différence entre deux traitements inférieure à LSD est non-significative.

Cette relation relie LSD et n !!!

Détermination du nombre de d'observations

On peut alors déterminer le nombre d'observations :

1. On se fixe la plus petite différence (LSD) que l'on veut pouvoir détecter (e.g. 10)
2. On choisit le seuil α
3. On cherche une valeur raisonnable de σ^2
4. On cherche la valeur de n pour laquelle le LSD est le plus proche de la valeur 1.

Exemple:

Calculation of SED and LSD for a CRD with $t = 4$ Treatments, Varying Replication (n) and Estimated Residual Variance $s^2 = 75$ (Example 10.1A)

Replication (n)	Units ($N = n \times t$)	Residual df ($N - t$)	$t_{N-t}^{[0.025]}$	SED	LSD
2	8	4	2.776	8.66	24.04
3	12	8	2.306	7.07	16.31
4	16	12	2.179	6.12	13.34
5	20	16	2.120	5.48	11.61
6	24	20	2.086	5.00	10.43
7	28	24	2.064	4.63	9.55

En pratique

On a donc besoin

1. d'une estimation de $\hat{\sigma}^2 = S_Y^2$
2. d'une idée de la différence que l'on veut pouvoir détecter

Il y a différentes manières d'obtenir cette information…

1. on peut réaliser une expérience préliminaire
2. la littérature peut nous donner des valeurs d'expériences précédentes
3. l'expertise du laboratoire
4. Un intervalle de $+/-2S$ comprend 95% des observations → $(max\text{-}min)/4$

Rq. Le même raisonnement s'applique dans le cas d'un modèle mixte

MAIS

1. la présence de termes de covariance non nuls entre certaines observations fait qu'il faut modifier la définition de SED (cf Ch 03)
2. le fait que SED n'est plus construit sur le MSE fait que le nombre de degrés de liberté doit, lui aussi, être revu

Puissance d'un test

La méthode présentée ci-dessus mène au nombre réplicats qui permet de rejeter H_0 pour une différence observée $\hat{\mu}_2 - \hat{\mu}_1$ avec une probabilité de se tromper de $1 - \alpha$.

Dans les faits, on est parfois davantage intéressé de maximiser la probabilité de détecter une différence $\mu_2 - \mu_1$ entre les vraies moyennes.

La question est alors de déterminer le nombre de réplicats qui minimise le risque β de ne pas rejeter H_0 alors qu'une différence existe ?

Petit rappel de proba-stat:

Terminology for Inferential Errors and Probabilities Associated with a Hypothesis Test

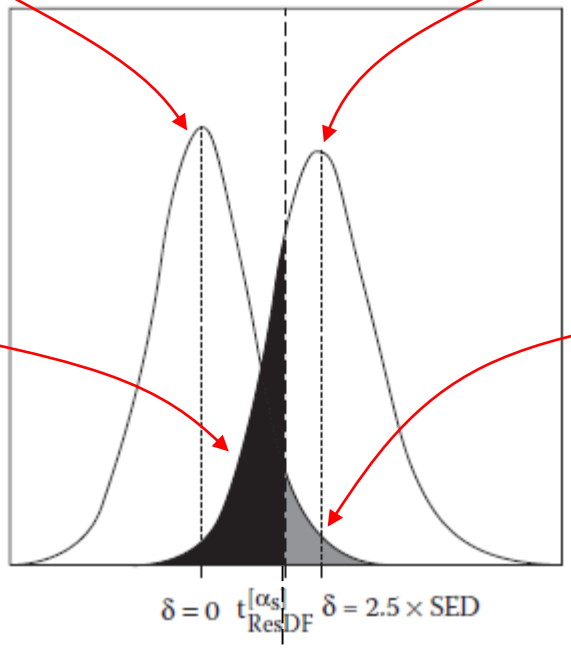
		Decision (Probability)	
		Accept H_0	**Reject H_0**
Null hypothesis (H_0)	**True**	Correct decision ($1 - \alpha_s$)	Incorrect decision (Type I error, α_s)
	False	Incorrect decision (Type II error, β_s)	Correct decision (Power, $1 - \beta_s$)

Puissance d'un test - représentation graphique

Distribution de
$t = (\hat{\mu}_2 - \hat{\mu}_1)/SED$
sous
H0: $\mu_2 - \mu_1 = 0$

Distribution de
$t = (\hat{\mu}_2 - \hat{\mu}_1)/SED$
sous
H1: $\mu_2 - \mu_1 = \delta$

Non-rejet à tort
H1 est vraie,
t inférieure au
seuil critique

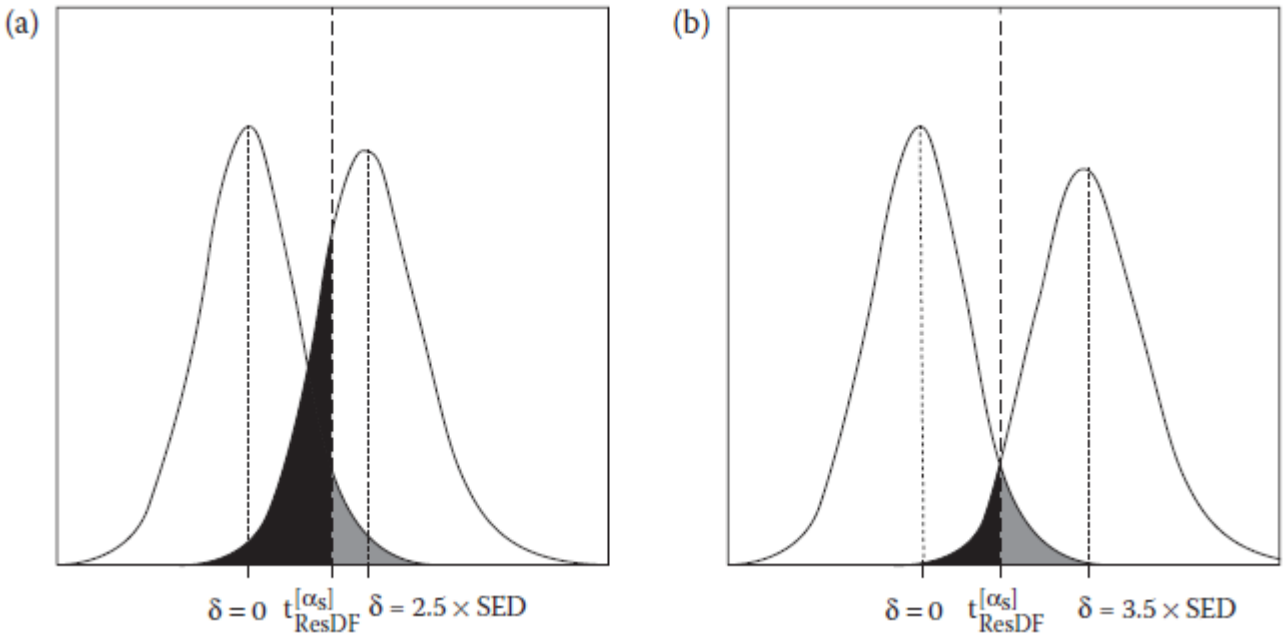


Rejet à tort
H0 est vraie,
t supérieure
au seuil
critique

Zone de non-rejet de H0
$|t| \leq t_{N-m\,;\,1-\alpha/2}$

Zone de rejet de H0
$|t| > t_{N-m\,;\,1-\alpha/2}$

Facteurs qui influencent la puissance d'un test



- SED (donc S_Y^2 et n), car S_Y^2 et n déterminent la forme des courbes H0 et H1
- la valeur critique de la statistique de test (donc n et α)
- δ, qui détermine la courbe H1

Détermination de la puissance d'un plan

On trouve plusieurs méthodes dans la littérature (et les logiciels) pour déterminer la puissance d'un plan donné. En « testant » une série de plans, on peut alors choisir le nombre de réplicats fournissant une puissance donnée.

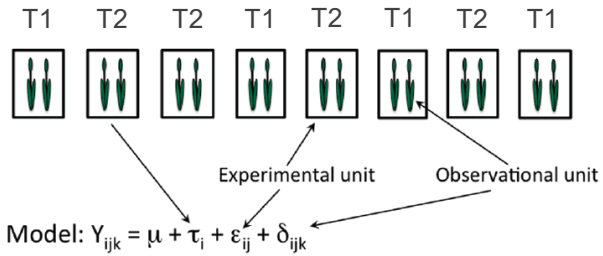
1. Si on peut écrire la distribution de la statistique de test sous l'hypothèse H1, alors on peut déterminer l'aire « noire » sous la courbe correspondant à H1 et donc la probabilité de ne pas rejeter H0 sous l'hypothèse H1.

2. Sinon, on peut simuler un grand nombre de jeux de données indépendants pour lesquels on pose la différence vraie entre les moyennes et réaliser le test pour chacun de ces jeux de données. La puissance du plan est alors approchée par la fraction des jeux de données pour lesquels le test a bien mené au rejet de H0. Cette méthode s'applique aisément à une grande variété de plans.

Gbur et al. (2012) Analysis of Generalized Linear Mixed Models in the Agricultural and Natural Resources Sciences, ASA-SSSA-CSSA publishing. (Chapitre 7)

Détermination de la puissance d'un plan

L'exemple ci-contre illustre l'influence du nombre x d'unités expérimentales et du nombre s de mesures par unité sur la puissance d'un plan comprenant 2 traitements, si $\mu_1 = 95, \mu_2 = 100, \sigma^2_{rep} = 5$ et $\sigma^2 = 10$



$s = 2$
$x = 4$

ANOVA Source of variation	df
Treatments	1 (fixed)
Experimental error	6
Observational error	8

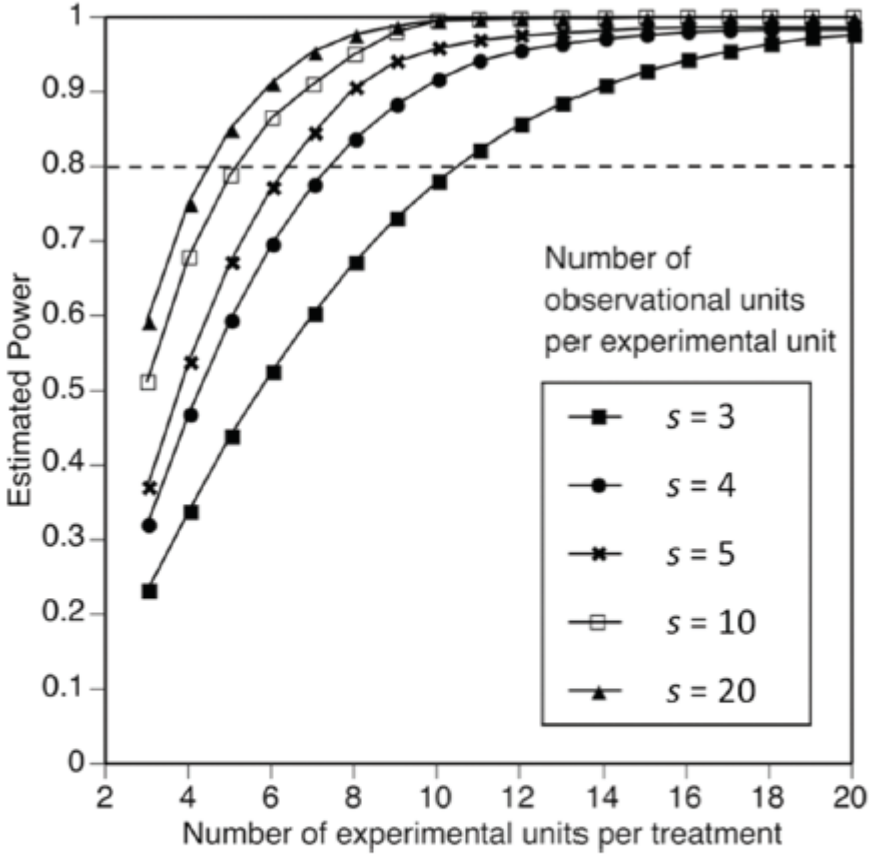


Fig. 5. Estimated power of a hypothesis test designed to detect a treatment difference of 5% of the mean with a Type I error rate of 0.05, variance component estimates of 5 and 10 (experimental and sampling errors, respectively), and varying numbers of experimental units and observational units (s = 3–20). The dashed line represents power = 0.8 and illustrates that different replication and sampling scenarios can be created to achieve the same result.

Agronomy Journal 107:692

Exemple en R avec le package simr (méthode 2: simulation)

```
require(lme4)
require(simr)
simrOptions(list(progress = FALSE))

## Etape 1: on crée un dataframe contenant les facteurs en colonne, et dont le nombre de lignes
     correspond au plan. Ce dataframe ne comprend pas la variable observée. Dans l'exemple qui suit, on
     considère deux traitements, quatre parcelles pour chaque traitement, une observation par parcelle.
X <- expand.grid(obs = 1, trt = c(1, 2), rep = 1:4)
X$trt <- as.factor(X$trt)
X$rep <- as.factor(X$rep)

## Etape 2: on crée les vecteurs des paramètres fixes (b) et aléatoires (V et s)
b <- c (90, 5) ## (m + alpha1, alpha2 - alpha1)
V <- 5         ## Variance rep
s <- sqrt(10)  ## erreur standard résiduelle

## Etape 3: on génère le modèle avec les paramètres forcés
mod1 <- makeLmer(y ~ trt + (1|rep:trt), fixef = b, VarCorr = V, sigma = s, data = X)
print(mod1)

Linear mixed model fit by REML ['lmerMod']
Formula: y ~ trt + (1 | rep)
   Data: X
REML criterion at convergence: 41.07
Random effects:
 Groups   Name        Std.Dev.
 rep      (Intercept) 2.236
 Residual             3.162
Number of obs: 8, groups:  rep, 4
Fixed Effects:
(Intercept)         trt2
         90            5
```

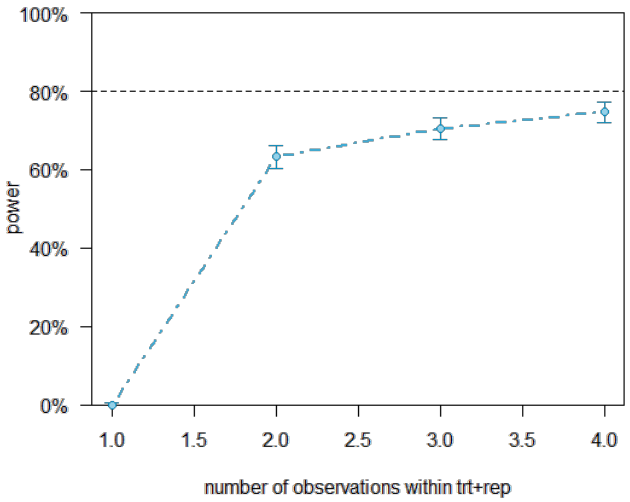
Exemple en R avec le package simr (méthode 2: simulation)

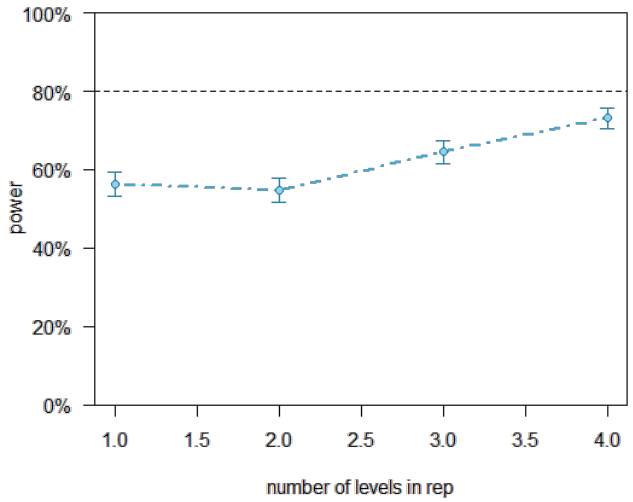
```
## Etape 4: on simule les jeux de données et calcule la puissance
simr::powerSim(mod1, nsim = 200)
```

```
Power for predictor 'trt', (95% confidence interval):
      80.00% (73.78, 85.31)

Test: Likelihood ratio

Based on 200 simulations, (1 warning, 0 errors)
alpha = 0.05, nrow = 8
```

Exemple en R avec le package simr (méthode 2: simulation)

```
## On crée le dataframe avec le nombre max de parcelles par traitement (rep) et d'observations par parcelle
X <- expand.grid(obs = 1:4, trt = c(1, 2), rep = 1:4)
X$trt <- as.factor(X$trt)
X$rep <- as.factor(X$rep)
b <- c (95, 5)
V <- 5
s <- sqrt(10)
mod1 <- makeLmer(y ~ trt + (1|rep:trt), fixef = b, VarCorr = V, sigma = s, data = X)

# On évalue la puissance pour un nombre variable d'observations
pc_obs <- powerCurve(mod1, within = 'trt+rep', breaks = 1:4, nsim = 1000)
plot(pc_obs)

# On évalue la puissance pour un nombre variable de rep
pc_rep <- powerCurve(mod1, along = 'rep', breaks = 1:4, nsim = 1000)
plot(pc_rep)
```





Différents niveaux de réplication

La réplication sert à estimer les différentes sources d'erreurs dans une expérience:

- variabilité entre différents échantillons d'une même unité expérimentale
- variabilité entre différentes unités expérimentales ayant le même traitement
- variabilité entre différents groupes (année, site, zone, expérimentateur, machine,…) d'unités expérimentales.

La réplication peut se réaliser à différents niveaux:

- plusieurs échantillons dans une unité expérimentale (pseudo-réplication)
- plusieurs unités expérimentales ayant le même traitement
- plusieurs groupes d'unités expérimentales
- …

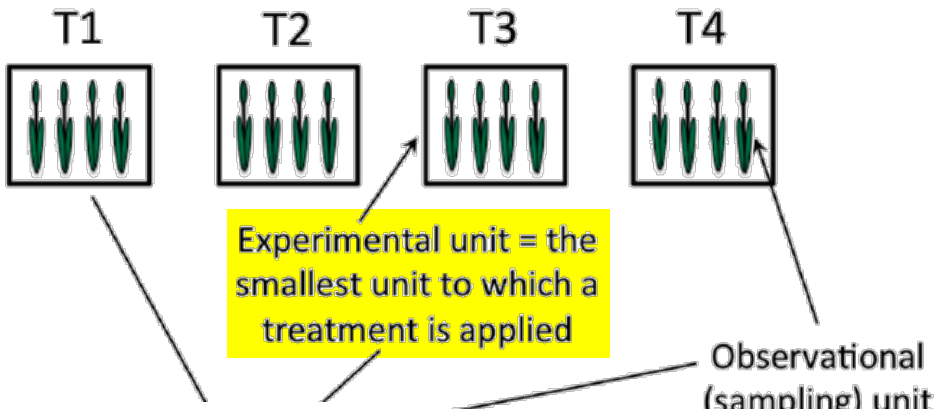
Ces différents niveaux de réplication et leurs conséquences sont illustrés dans les dias suivantes, avec différentes expériences dans lesquelles 16 parcelles sont utilisées pour analyser les effets de 4 traitements.

Un peu de vocabulaire pour parler le même language

Term	Definition
Experiment	a planned and organized inquiry designed to test a hypothesis, answer a question, or discover new facts
Treatment	a procedure or system whose effect on the experimental material is to be measured or observed
Experimental unit	the smallest unit to which a treatment is applied
Observational unit	the unit upon which observations or measurements are made
Block	a group of (presumably) homogeneous experimental units (a complete block contains all treatments)
Experimental design	the set of rules and procedures by which the treatments are assigned to experimental units
Treatment design	the organization or structure that exists across the treatments used to define the experiment
Replication	the practice of applying each treatment to multiple and mutually independent experimental units
Randomization	the practice of assigning treatments to experimental units such that each unit is equally likely to receive each treatment
Factor	a type of treatment; this can take on many forms: quantitative, qualitative, ranked, or nested
Level	a specific form or "state" of a factor
Factorial treatment	a combination of one level of each factor used to create a unique treatment combination
Experimental error	the variance among experimental units treated alike, often symbolized as σ^2 or σ_e^2.
Sampling error	the variance among observational units within experimental units; there can be multiple levels of sampling error
Precision	the inverse of experimental error, $1/\sigma_e^2$
Confounding	the purposeful or inadvertent mixing of two or more effects, such that no statistical analysis can separate them

Agronomy Journal 107:692

Différents niveaux de réplication



Model: $Y_{ij} = \mu + \alpha_i + \varepsilon_i + o_{ij}$

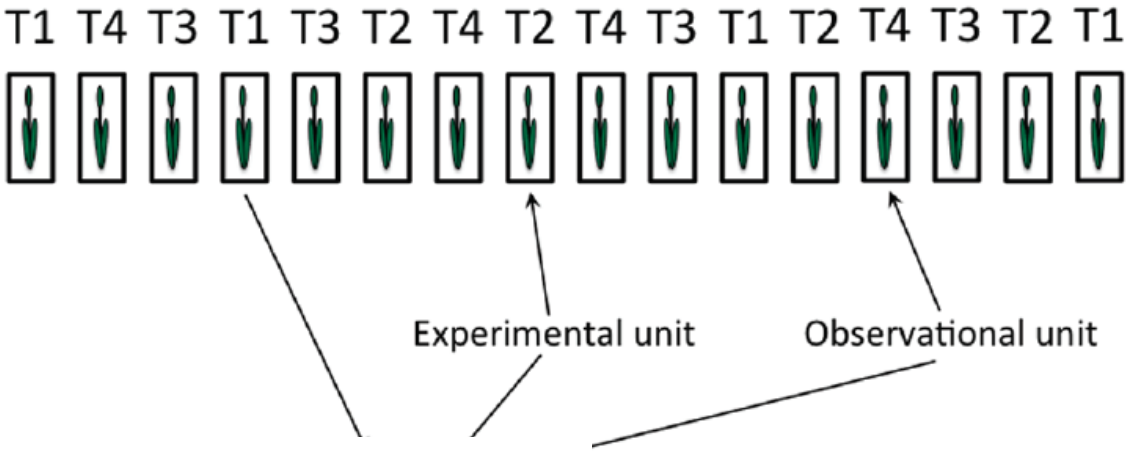
ANOVA Source of variation	df
Treatments + Experimental error	3 (fixed)
Observational error	12

Traitement et unité expérimentale sont confondus.

L'effet fixe des traitements contient une composante aléatoire dont la grandeur ne peut pas être estimée. On ne peut donc pas vraiment attribuer l'effet traitement aux traitements appliqués.

La seule composante aléatoire estimée est la résiduelle et elle est souvent plus faible que la composante expérimentale. Le test F risque donc d'être plus élevé.

Différents niveaux de réplication



Completely randomized design (CRD)

Model: $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} + o_{ij}$

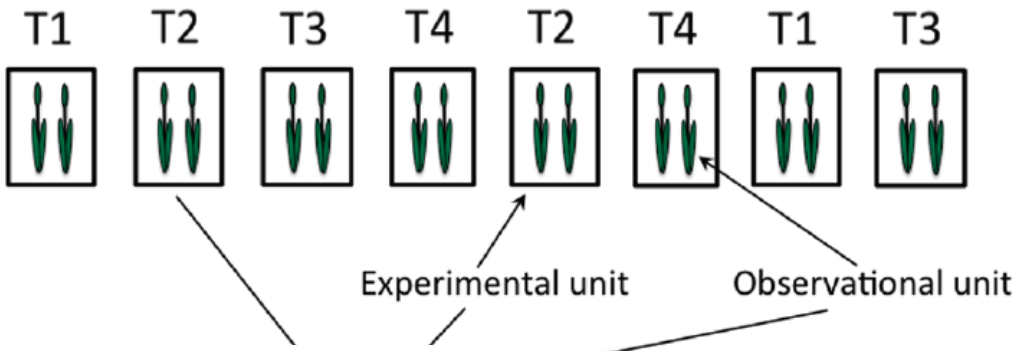
ANOVA Source of variation	df
Treatments	3 (fixed)
Error (experimental + observational)	12

Unité expérimentale et unité d'observation sont confondus.

Avec ce nombre d'unités expérimentales, les traitements sont bien répliqués.

La confusion entre les unités expérimentale et d'observation fait que l'on ne peut les estimer séparément.

Différents niveaux de réplication



Completely randomized design (CRD)

Model: $Y_{ijk} = \mu + \alpha_i + \varepsilon_{ij} + o_{ijk}$

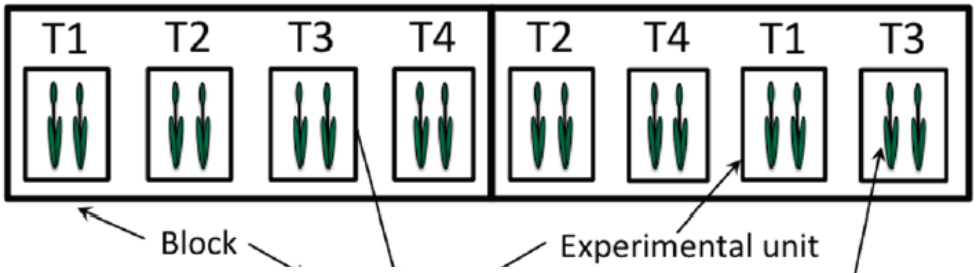
ANOVA Source of variation	df
Treatments	3 (fixed)
Experimental error	4
Observational error	8

On a ici deux niveaux de réplication: deux unités expérimentales au niveau des traitements et deux unités d'observation au sein de chaque unité exp.

On peut ici estimer séparément les deux composantes de variance.

Les unités d'observation d'une même unité exp. ne sont pas indépendantes.

Différents niveaux de réplication



Model: $Y_{ijk} = \mu + b_j + \alpha_i + \varepsilon_{ij} + o_{ijk}$ — Observational unit

ANOVA Source of variation	df
Blocks	1
Treatments	3 (fixed)
Experimental error	3
Observational error	8

Randomized complete block design (RCBD)

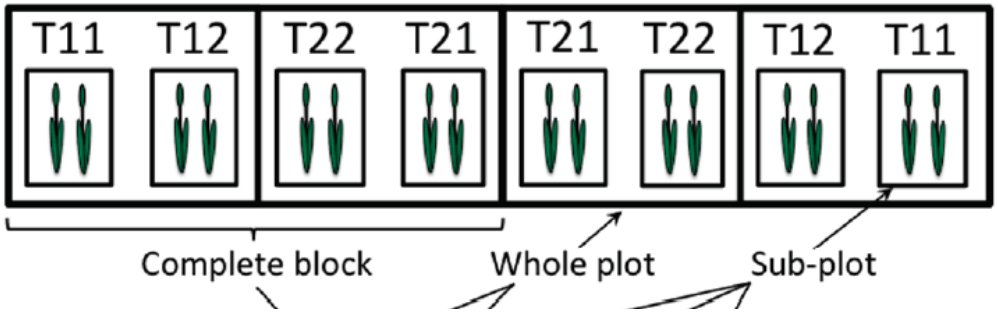
Les unités expérimentales sont groupées en blocs, chaque bloc comprenant l'ensemble des traitements. Ceci permet de prendre en compte les effets communs aux unités expérimentales d'un même bloc (e.g. exposition serre)

On peut estimer séparément les deux composantes de variance et l'effet bloc

Les unités d'observation d'une même unité exp. ne sont pas indépendantes

Les unités expérimentales d'un même bloc ne sont pas indépendantes

Le cas du split-plot design (pour réfléchir)



Model $Y_{ijkl} = \mu + b_k + \alpha_i + o_{ik} + \beta_j + \alpha\beta_{ij} + e_{ijk} + o_{ijkl}$

ANOVA Source of variation	df
Blocks	1
Factor A	1 (fixed)
Whole plot error	1
Factor B	1 (fixed)
A x B Interaction	1 (fixed)
Sub-plot error	2
Observational error	8

Deux facteurs fixes (A et B) croisés, deux niveaux chacun.

Deux blocs complets (on y trouve les 4 traitements)

Au sein des blocs, on a deux sous-blocs, confondus avec le facteurs A

Quelques leçons

Les réplications sont nécessaires pour quantifier les sources d'erreur

La réplication peut être réalisée à différents niveaux

- Elle doit au moins l'être au niveau de l'unité expérimentale (l'unité la plus petite à laquelle on applique un traitement)
- Elle doit l'être à des échelles spatiales et temporelles qui correspondent avec l'application des traitements

Déterminer le nombre de réplication n'est pas qu'une question de feeling, de protocole, ou d'imitation.

On peut déterminer le nombre de réplications en fonction de plusieurs cibles (nous en avons évoqué deux)

Déterminer le nombre de réplications est une décision essentielle qui conditionne largement la valeur de l'expérience. Les chapitres suivants traitent d'autres décisions…

Glâné dans la littérature…

(Cochran and Cox, 1992), "It has come to be recognized that the time to think about statistical inference, or to seek [a statistician's] advice, is when the experiment is being planned." Hahn (1984) put it more forcefully, "Statisticians make their most valuable contributions if they are consulted in the planning stages of an investigation. Proper experimental design is often more important than sophisticated statistical analysis." He continues, quoting H. Ginsburg as saying, "When I'm called in after it's all over, I often feel like a coroner. I can sign the death certificate—but do little more." Light et al. (1990) stated it slightly differently, "You cannot save by analysis what you bungle by design."

Gbur et al. (2012)

A failed experimental design is generally manifested as an experiment with high *P* values, leaving the researcher with uncertain or equivocal conclusions: are the treatments really not different from each other, is my experimental design faulty due to poor planning and decision making, or was there some unknown and unseen disturbance that occurred to the experiment, causing errors to be inflated? Rarely can these questions be answered when *P* values are high, resulting in unpublishable results and wasted time and money.

Agronomy Journal 107:692