NOMBRE DE SUJETS NÉCESSAIRES (NSN)

– LE CAS DES ÉTUDES OBSERVATIONNELLES

MARIE ANSOBORLO
EMELINE LAURENT
MARC TASSI

PlaMeth
03/09/2024

OBJECTIF DE CETTE PLAMETH

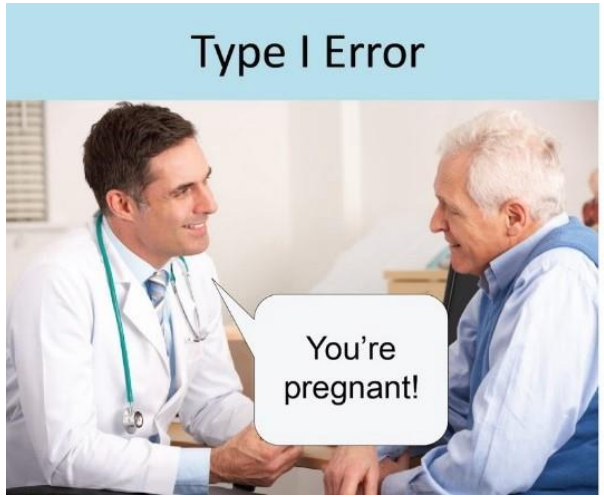
- Savoir répondre à ce type de demande en CoMeth ou contexte similaire :

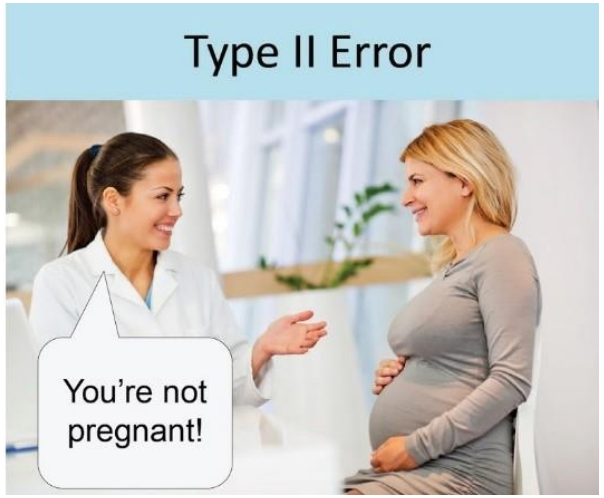
Quel est le nombre de sujets nécessaires dans mon étude (observationnelle) ?
NB : je soutiens ma thèse dans 6 mois

AVANT-PROPOS : RISQUES D'ERREUR

- risque α : risque de rejeter l'hypothèse nulle à tort
- risque β : risque de ne pas rejeter l'hypothèse nulle à tort

		Décision	
		Rejeter H0	Ne pas rejeter H0
Réalité	H0 vraie	α	1-α
	H1 vraie	1-β	β





1- β = puissance

- Aptitude à mettre en évidence un effet s'il existe
- Calculable a priori sur la base d'hypothèses

QUEL LIEN AVEC LE NSN?

Le NSN porte sur le risque β (habituellement fixé à 20 %) et non sur le risque α (à 5 %)

NSN = Plus petit effectif théorique qui permettra de garantir une puissance donnée, c'est-à-dire une probabilité de mettre en évidence un effet s'il existe.

Cela veut donc dire qu'en incluant un nombre de sujets = NSN, avec une puissance de 80 % il existe encore 20 % de risque de ne pas mettre en évidence une différence alors qu'elle existe.

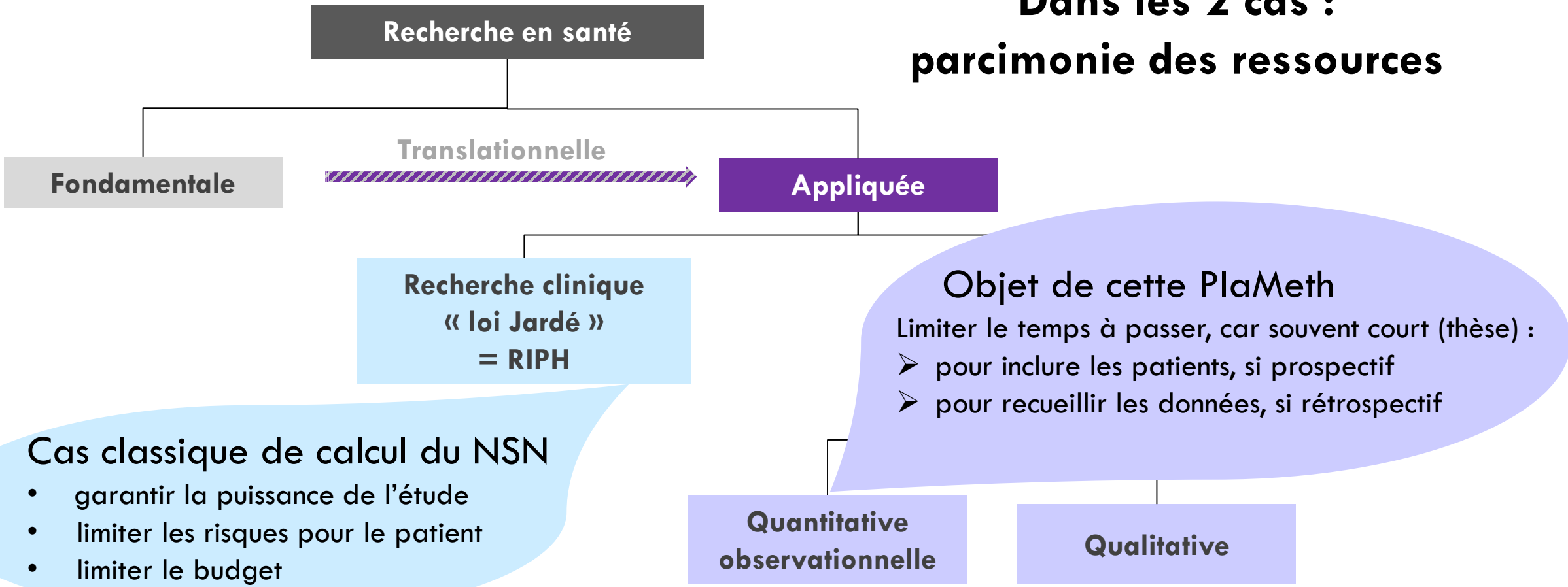
Attention, ce n'est valable que pour le critère de jugement principal +++

Autrement dit, si des analyses en sous-groupes sont prévues, la puissance risque de ne plus être suffisante et il faut faire des calculs pour tous les effectifs et choisir l'effectif le plus défavorable !!!

Et ça demande de faire beaucoup d'hypothèses, issues de la littérature le plus souvent

Travail bibliographique en amont par les demandeurs +++

UN NSN, POURQUOI ?



Dans les 2 cas :
parcimonie des ressources

Cas classique de calcul du NSN
- garantir la puissance de l'étude
- limiter les risques pour le patient
- limiter le budget

Objet de cette PlaMeth

Limiter le temps à passer, car souvent court (thèse) :
- pour inclure les patients, si prospectif
- pour recueillir les données, si rétrospectif

UN NSN, POURQUOI ?

Corollaire (important !)

Si inclure 1 000 000 de sujets vs 100 sujets requiert le <u>même effort</u>, la question du NSN ne se pose pas

Ex : études rétrospectives sur bases de données médico-administratives (BDMA : PMSI, SNDS,…)

(ni celle du test statistique, potentiellement…mais cela est une autre histoire)

Objet de cette PlaMeth

Limiter le temps à passer, car souvent court (thèse) :

- pour inclure les patients, si prospectif
- pour recueillir les données, si rétrospectif

Quantitative observationnelle

LA PREMIERE QUESTION A POSER

Quel est ton objectif ?

Je veux une estimation suffisamment précise de ma moyenne/ mon pourcentage

Je veux montrer une différence significative…

…entre 2 groupes

…par rapport à une valeur théorique (littérature)

Ex : « J'essaie de monter une étude de phase 2 sur la radiothérapie dans les tumeurs de vessie non infiltrantes. L'objectif est la survie sans récidive. Ils me demandent le traitement standard qui est la BCG thérapie avec une survie de 90%. Et l'effet attendu de la radiothérapie qui est de 80% dans les études."

ET ENSUITE

Combien de temps as-tu pour faire ton étude ?

Combien de sujets penses-tu pouvoir inclure sur ce temps-là ?

Euh…dans les 2 cas : pas beaucoup

1bis.
Pour vérifier la faisabilité (ou infaisabilité) de ton étude, passe en diapo 10

1.
Si tu t'attends à peu d'effectif, il y a très peu de chance que tu arrives à montrer une différence significative (ou obtenir une estimation suffisamment précise) ➔ **reconsidérer l'objectif ?**

2.
Si tu souhaites malgré tout mener l'étude, inclus le maximum possible, avec objectif purement descriptif

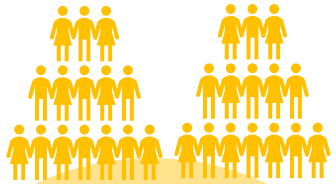
2bis.
Ça, quel que soit le cas, ce ne sera jamais faux !

Et prévois au moins 1-2 mois pour data-management (dont qualité) + analyse + rédaction

ET ENSUITE

Combien de temps as-tu pour faire ton étude ?

Combien de sujets penses-tu pouvoir inclure sur ce temps-là ?



J'ai tout le temps et les effectifs du monde, mais je veux optimiser

Ok, on passe à l'étape suivante

C'est l'heure de démontrer mes super-pouvoirs !!

Je veux une estimation suffisamment précise de ma moyenne/ mon pourcentage

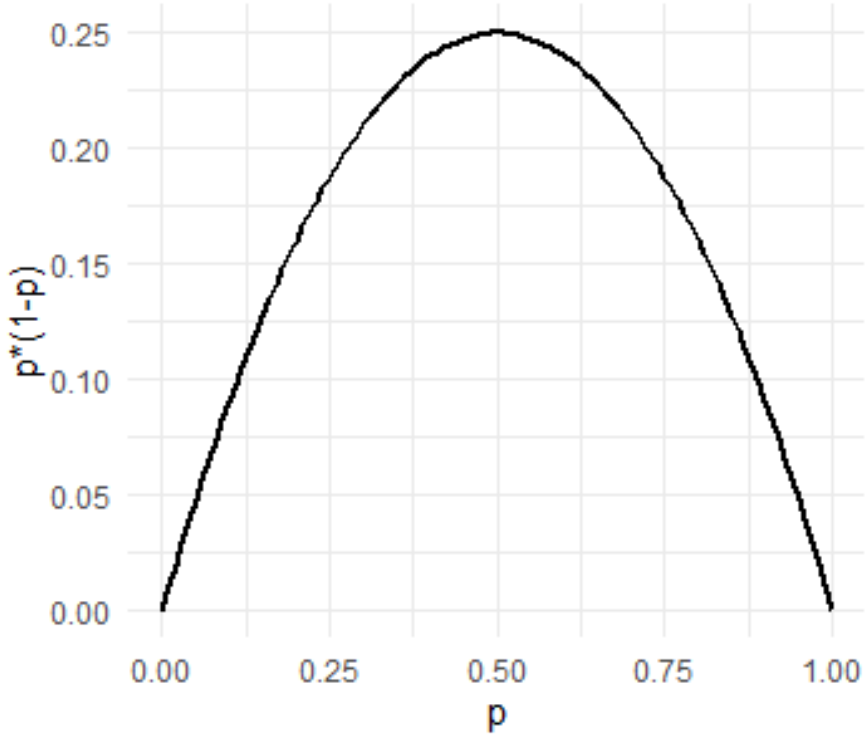
D'accord,
tu veux un intervalle de confiance

Outil eCDC :
Portail Recherche > Nombre de sujets pour une étude descriptive (en bas de page)

ESTIMER LA LARGEUR DE L'INTERVALLE DE CONFIANCE

Plus p est proche de 50 %, plus l'intervalle de confiance sera large, donc plus il faudra inclure de patients pour augmenter la précision

$$p \pm z_{\alpha} \times \sqrt{\frac{p(1-p)}{n}}$$



Je veux une estimation suffisamment précise de ma moyenne/ mon pourcentage

D'accord,
tu veux un intervalle de confiance

Outil eCDC :
Portail Recherche > Nombre de sujets pour une étude descriptive (en bas de page)

Je veux montrer une différence significative…

…entre 2 groupes

…par rapport à une valeur théorique (littérature)

EasyMedStat :
https://www.easymedstat.com/fr/calculateur-taille-echantillon

J'ai obtenu mon nombre de sujets nécessaires. Super je me lance !

Une dernière chose : inclus 20-25 % de sujets en plus (déjà pris en compte par EasyMedStat), afin de tenir compte des sujets que tu devras exclure suite à ton recueil de données : données manquantes ou aberrantes

ET CE SERA FIABLE A TOUS LES COUPS ?

Oui si toutes les hypothèses sont respectées (ça fait beaucoup !)

Variable suivant le test utilisé pour les analyses

A titre indicatif (non exhaustif suivant l'analyse effectuée) :

Proportion	Moyenne
Descriptif	
La proportion estimée p est exacte	La moyenne estimée est exacte La variance estimée est exacte La variable suit une distribution normale
Les observations sont mutuellement indépendantes La probabilité p de posséder la propriété est la même pour toutes les observations L'effectif N est assez grand, et p n'est pas trop proche de 0 ou de 1.	
Analytique (comparaison)	
Si comparaison de 2 groupes, toutes les hypothèses doivent être validées sur chacun des 2 groupes	
	La variance des deux groupes est identique

EXEMPLES D'OUTILS

EasyMedStat.com (no code, no stat friendly, pas à pas,…)

D'autres outils existent ! … gratuits …

BiostaTGV.sentiweb.fr

- https://cran.r-project.org/web/packages/epiR/epiR.pdf

➔ Autant d'outils que d'estimations différentes ?!

https://homepage.univie.ac.at/robin.ristl/samplesize.php

VARIABILITÉ INTER-LOGICIELS/MÉTHODES

ELSEVIER

Journal of Clinical Epidemiology 67 (2014) 601–605

Journal of Clinical Epidemiology

A myriad of methods: Calculated sample size for two proportions was dependent on the choice of sample size formula and software

Melanie L. Bell[a,b], Armando Teixeira-Pinto[c,*], Joanne E. McKenzie[d], Jake Olivier[e]

[a]*Psycho-Oncology Co-Operative Research Group, School of Psychology, University of Sydney, Australia*
[b]*Mel and Enid Zuckerman College of Public Health, University of Arizona, 295 N Martin Ave, Tucson, AZ 85724, USA*
[c]*School of Public Health, Edward Ford Building (A27), University of Sydney, Sydney NSW 2206, Australia*
[d]*School of Public Health and Preventive Medicine, Alfred Centre, Monash University, Melbourne VIC 3004, Australia*
[e]*School of Mathematics and Statistics, The Red Centre, The University of New South Wales, Sydney 2052, Australia*

Accepted 21 October 2013; Published online 16 January 2014

A myriad of methods: Calculated sample size for two proportions was dependent on the choice of sample size formula and software - ScienceDirect

epi.sssimpleestb	Sample size to estimate a binary outcome using simple random sampling.
epi.sssimpleestc	Sample size to estimate a continuous outcome using simple random sampling.
epi.ssstrataestb	Sample size to estimate a binary outcome using stratified random sampling.
epi.ssstrataestc	Sample size to estimate a continuous outcome using stratified random sampling.
epi.ssclus1estb	Sample size to estimate a binary outcome using one-stage cluster sampling.
epi.ssclus1estc	Sample size to estimate a continuous outcome using one-stage cluster sampling.
epi.ssclus2estb	Sample size to estimate a binary outcome using two-stage cluster sampling.
epi.ssclus2estc	Sample size to estimate a continuous outcome using two-stage cluster sampling.
epi.ssxsectn	Sample size, power or detectable prevalence ratio for a cross-sectional study.
epi.sscohortc	Sample size, power or detectable risk ratio for a cohort study using count data.
epi.sscohortt	Sample size, power or detectable risk ratio for a cohort study using time at risk data.
epi.sscc	Sample size, power or detectable odds ratio for case-control studies.
epi.sscompb	Sample size, power and detectable risk ratio when comparing binary outcomes.
epi.sscompc	Sample size, power and detectable risk ratio when comparing continuous outcomes.
epi.sscomps	Sample size, power and detectable hazard when comparing time to event.
epi.ssequb	Sample size for a parallel equivalence or equality trial, binary outcome.
epi.ssequc	Sample size for a parallel equivalence or equality trial, continuous outcome.
epi.sssupb	Sample size for a parallel superiority trial, binary outcome.
epi.sssupc	Sample size for a parallel superiority trial, continuous outcome.
epi.ssninfb	Sample size for a non-inferiority trial, binary outcome.
epi.ssninfc	Sample size for a non-inferiority trial, continuous outcome.
epi.ssdetect	Sample size to detect an event.
epi.ssdxsesp	Sample size to estimate the sensitivity or specificity of a diagnostic test.
epi.ssdxtest	Sample size to validate a diagnostic test in the absence of a gold standard.

T-TEST / STUDENT POUR COMPARER 2 MOYENNES OBSERVÉES (2 ÉCHANTILLONS)

H0 : m (groupe A) = m (groupe B)

Alpha 5%, Puissance 80%

m (A)	**m (B)**	**ET**	**delta**	**N**
2.1	2.2	0.5	0.1	?

T-TEST / STUDENT POUR COMPARER 2 MOYENNES OBSERVÉES (2 ÉCHANTILLONS)

H0 : m (groupe A) = m (groupe B)

m (A)	**m (B)**	**ET**	**delta**	**N**
2.1	2.2	0.5	0.1	786

APPLICATION NUMÉRIQUE

Le NSN pour rejeter l'hypothèse nulle, augmente lorsque ...

m (A)	m (B)	ET	delta	N
2.1	2.2	0.5	0.1	786
4.2	4.7	0.5	0.5	32

... la taille d'effet que l'on souhaite mettre en évidence diminue

(peu importe la valeur absolue des paramètres étudiés)

Fleiss, Statistical Methods for Rates and Proportions, formulas 3.18 &3.19
(Fleiss JL. 1981. pp. 44-45)

APPLICATION NUMÉRIQUE

Le NSN pour rejeter l'hypothèse nulle, augmente lorsque ...

m (A)	m (B)	ET	delta	N
2.1	2.2	0.5	0.1	786
4.2	4.7	0.5	0.5	32
6.03	6.08	0.5	0.05	3 140

... la taille d'effet que l'on souhaite mettre en évidence diminue

(peu importe la valeur absolue des paramètres étudiés)

Fleiss, Statistical Methods for Rates and Proportions, formulas 3.18 &3.19
(Fleiss JL. 1981. pp. 44-45)

APPLICATION NUMÉRIQUE

Le NSN pour rejeter l'hypothèse nulle, augmente lorsque ...

m (A)	m (B)	ET	delta	N
2.1	2.2	0.7	0.1	1 540
4.2	4.3	0.4	0.1	504
6.03	6.136	0.9	0.1	2 544

... l'écart-type du paramètre que l'on étudie augmente

CHI-2 PEARSON POUR COMPARER 2 PROPORTIONS OBSERVÉES (2 ÉCHANTILLONS)

p (A)	**p (B)**	**N**
0,55	0,45	392
0,55	0,05	12
0,55	0,54	77 852

STROBE

STRENGTHENING
THE **R**EPORTING
OF **OB**SERVATIONAL STUDIES
IN **E**PIDEMIOLOGY

« Taille de l'étude –

Expliquer comment a été déterminé le nombre de sujets à inclure »

https://www.strobe-statement.org/

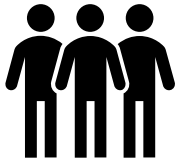
https://bookdown.org/melissaksharp/STROBE_eduexpansion/

STROBE Statement—checklist of items that should be included in reports of observational studies

	Item No	Recommendation
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found
Introduction		
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported
Objectives	3	State specific objectives, including any prespecified hypotheses
Methods		
Study design	4	Present key elements of study design early in the paper
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection
Participants	6	(*a*) *Cohort study*—Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up *Case-control study*—Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls *Cross-sectional study*—Give the eligibility criteria, and the sources and methods of selection of participants
		(*b*) *Cohort study*—For matched studies, give matching criteria and number of exposed and unexposed *Case-control study*—For matched studies, give matching criteria and the number of controls per case
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group
Bias	9	Describe any efforts to address potential sources of bias
Study size	10	Explain how the study size was arrived at
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding
		(*b*) Describe any methods used to examine subgroups and interactions
		(*c*) Explain how missing data were addressed
		(*d*) *Cohort study*—If applicable, explain how loss to follow-up was addressed *Case-control study*—If applicable, explain how matching of cases and controls was addressed *Cross-sectional study*—If applicable, describe analytical methods taking account of sampling strategy
		(*e*) Describe any sensitivity analyses

Continued on next page

POUR ALLER PLUS LOIN…



COMPARER DEUX MOYENNES

$$n = (Z_{\alpha} + Z_{2\beta})^2 \times 2 \times \sigma^2 / d^2$$

- Z_{α} = valeur de la distribution normale pour un α choisi (exemple : Z_{α} = 1,96 pour un niveau de confiance de 95%, α = 5% pour une hypothèse bilatérale*)
- $Z_{2\beta}$ = valeur de la distribution normale pour un β choisi (exemple : $Z_{2\beta}$ = 0,842 pour une puissance de 80 %, β est 20%)
- σ^2 = variance de la population
- d = différence que vous souhaitez détecter

NB : Il s'agit d'une formule basée sur la distribution normale

**cad une étude d'équivalence, H0 « différence entre les deux moyennes = 0 »*

DIFFÉRENCE ENTRE 2 PROPORTIONS

Intervalle de confiance à 95 %

$$(P_1 - P_2) \pm Z_\alpha \sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}$$

Mais ce n'est pas la seule méthode existante !! Ex : test du Chi-2

DIFFÉRENCE ENTRE 1 PROPORTION OBSERVÉE ET 1 PROPORTION THÉORIQUE

Intervalle de confiance à 95 %

$$p_0 \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{p_0\left(1-p_0\right)}{n}}$$

http://gandalfmagicien.free.fr/psycho/Licence%203/Premier%20Semestre/CM%20Maths/chapitre4.pdf

RÉFÉRENCES

https://www.infectiologie.com/UserFiles/File/formation/desc/2019/seminaire-avril-2019/jeudi-04-04-2019/recherche-8b-jeudi-04-nathanael-lapidus.pdf

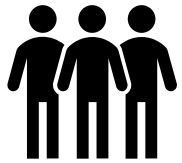
http://epivf.fr/calcul_nombre_sujets_intro.html

Correction de continuité :

https://influentialpoints.com/notes/n9cont.htm

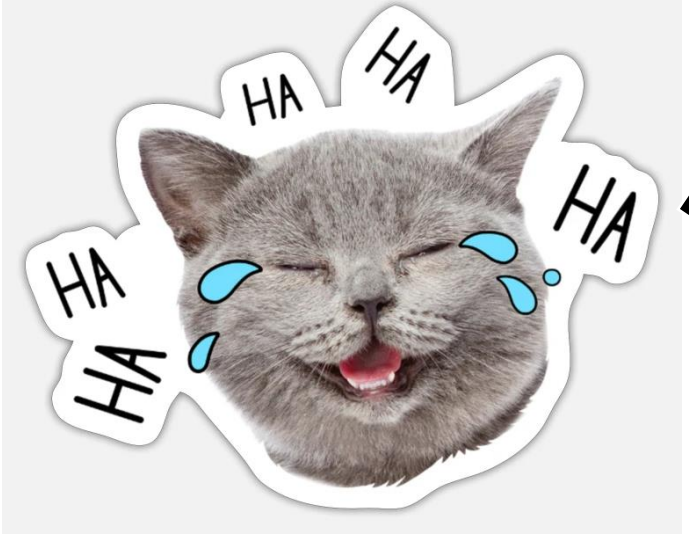
http://www.sthda.com/english/wiki/two-proportions-z-test-in-r

http://www.sthda.com/english/wiki/one-proportion-z-test-in-r

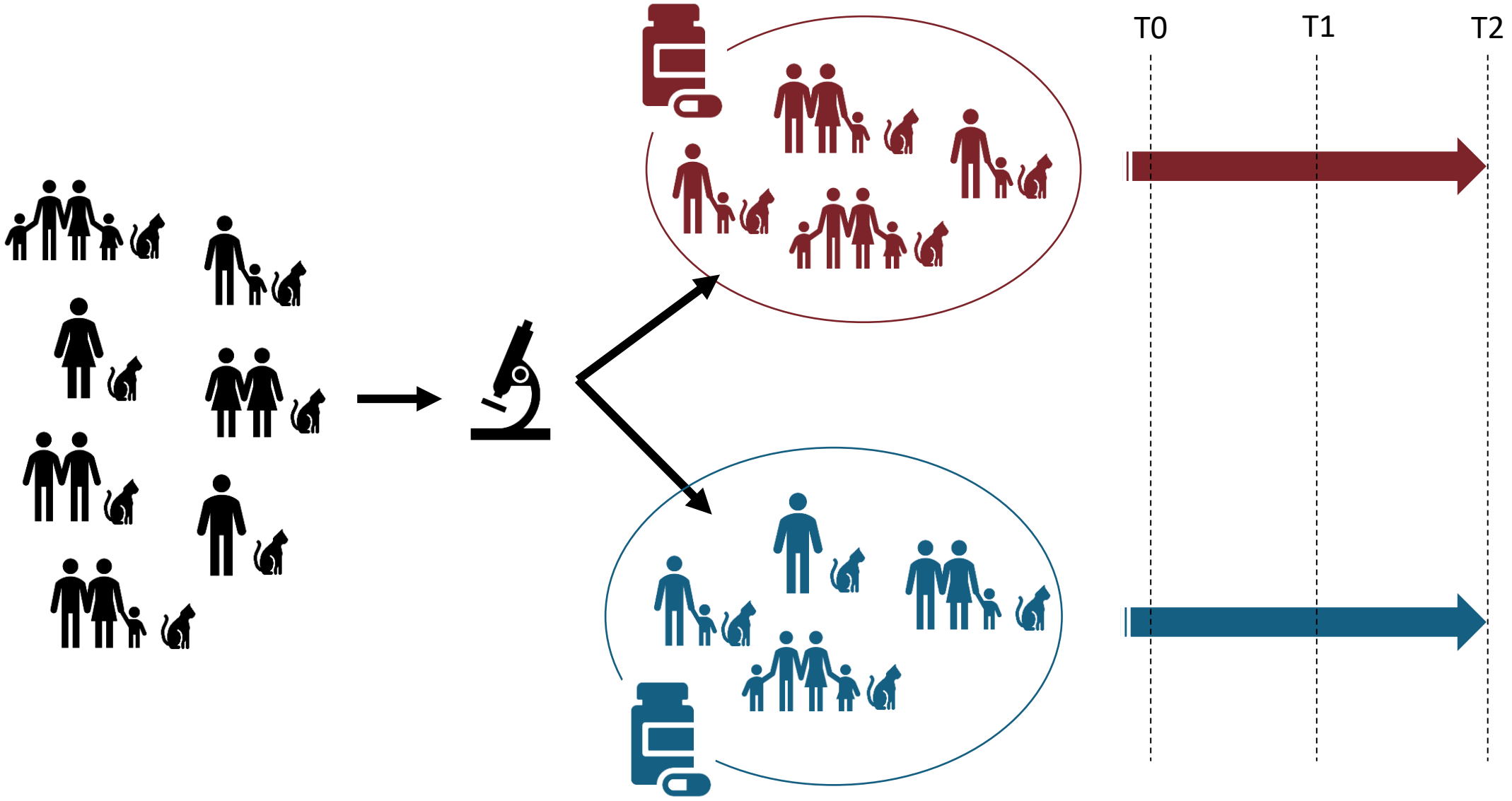


…ET ENCORE PLUS LOIN : EXEMPLE D'UNE ÉTUDE COMPLEXE

Contexte

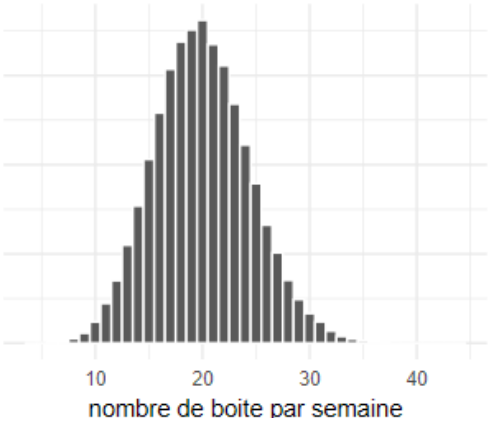


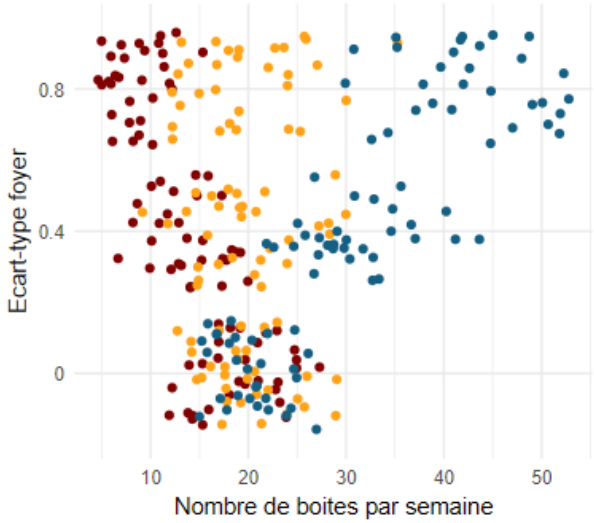
Design

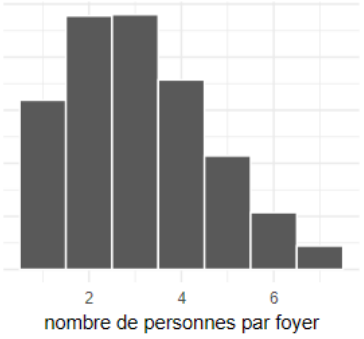


Hypothèses

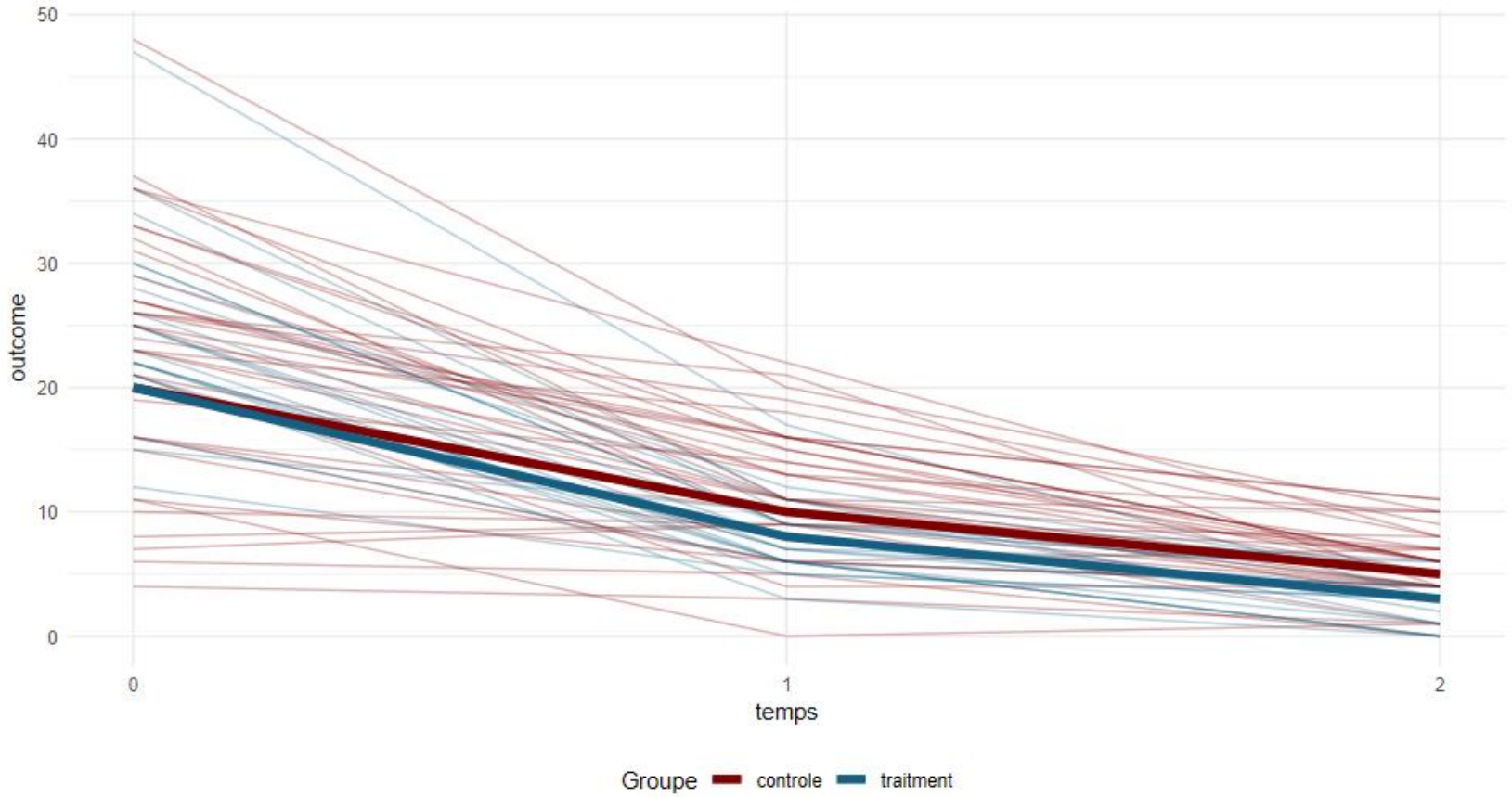
- Consommation à T0 → loi de Poisson et moyenne = 20 boites/semaine
- Effets à T1:
  - Groupe contrôle → réduction de la consommation de 40%
  - Groupe expérimental → réduction de la consommation de 50%
- Effets à T2:
  - Groupe contrôle → réduction de la consommation de 75%
  - Groupe expérimental → réduction de la consommation de 85%
- Effet famille : effet « aléatoire » : $N(0, 0.3)$
- Effet individu : effet « aléatoire » : $N(0, f)$; $f \sim Exp(5)$
- Nombre de personnes par foyer → loi de Poisson tronquée [1-7] de moyenne = 3







Hypothèses



Hypothèses

$$y \sim Poisson(\lambda)$$

Effet traitement — Effet temps — Interaction traitement / temps

$$\log y_{ijt} = \beta_{0ij} + \beta_M . Med + \beta_t . Tps + \beta_{MT} . Med . Tps + \varepsilon_{ijt}$$ → Résidu (déviation de la prédiction individuelle) — Niveau 1 (temps)

$$\beta_{0ij} = \gamma_{0j} + u_{ij}$$ → Intercept aléatoire individuel (déviation de la prédiction familiale) — Niveau 2 (individu)

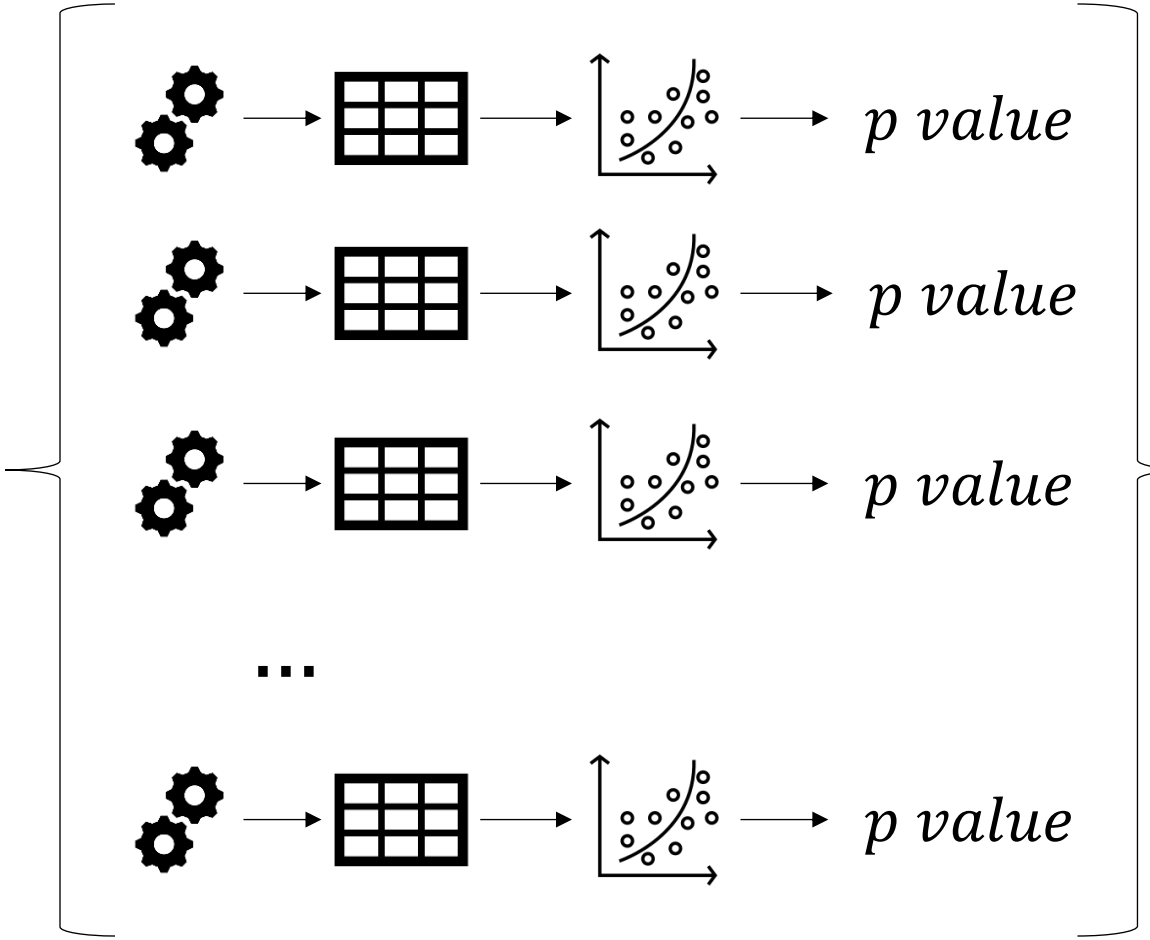
$$\gamma_{0j} = \delta_0 + v_j$$ → Intercept aléatoire familial (déviation de l'intercept fixe) — Niveau 3 (foyer)

Intercept fixe

$$\varepsilon_{ijt} \sim N(0, \sigma_e^2) \quad u_{ij} \sim N(0, \sigma_{uj}^2) \quad v_j \sim N(0, \sigma_v^2)$$

NSN ???

$N = n$



$$Pwr_{N=n} = \frac{\#\, p < 0{,}05}{\#\, echantillon}$$

Exemple avec R

```r
library(tidyverse)
library(extraDistr)
library(glmmTMB)
library(PropCIs)

# initialisation
res <- tibble(
  p.value = double(),
  nclust = double()
)

# nb de cluster (foyers) 6 à 30
for (i in seq(from = 6, to = 30, by = 2)) {

  nclust <- i

  # 200 simulations par taille de cluster
  for (j in 1:200) {

    # nombre de personnes par foyer -> distribution de poisson tronquée [0-7] de moyenne 3
    ind <- rtpois(n = nclust, lambda = 3, a = 0, b = 7)

     # la moitié des foyers ont le traitement de référence et l'autre moitié le traitement expériemental
    treat <- rep(0:1, nclust/2)

      # Effet famille
    uf = rnorm(n = nclust, mean = 0, sd = 0.3)

      # variance individus dans famille
    sd_f = rexp(n = nclust, rate = 5)
```

1

Exemple avec R

```r
 #Génération tableau
famille <-  tibble(
  # identifiant individu
  id = 1:sum(ind),
  # identifiant famille
  famille = (rep(1:nclust, ind)),
  # identifiant traitement
  treat = rep(treat, ind),
  # 20 prise par semaine à T0
  no = log(20),
  # Effet controle à T1
  t1 = log(0.5),
  # Effet controle à T2
  t2 = log(0.25),
  # Effet traitement à T1
  t1tr = log(0.4/0.5),
  # Effet traitement à T2
  t2tr = log(0.15/0.25),
  # Effet famille
  uf   = (rep(uf, ind))
  # Variance ind dans famille
  sd_f   = (rep(sd_f, ind))
)
```

2

```r
famille <- famille %>% mutate(
       # Effet individuel
       ui =  rnorm(n = n(), mean = 0, sd = sd_f),
       # Outcome à T0
       T0 = rpois(n = n(), lambda = exp(n0 + uf + ui)),
       # Outcome à T1
       T1 = case_when(
         treat == 0 ~ rpois(n = n(), lambda = exp(n0 + uf + ui + t1)),
         T ~     rpois(n = n(), lambda = exp(n0 + uf + ui + t1 + t1tr))
       ),
       # Outcome à T2
       T2 = case_when(
         treat == 0 ~rpois(n = n(), lambda = exp(n0 + uf + ui + t2)),
         T ~rpois(n = n(), lambda = exp(n0 + uf + ui + t2 + t2tr)),
       )
     ) %>%
     select(id, famille, treat, T0, T1, T2) %>%
     # Format de données long
     pivot_longer(cols = c(T0, T1, T2), names_to = 'temps', values_to = 'outcome')
```

3

Exemple avec R

4

```
    # Modèle linéaire généralisé avec effet aléatoire, distribution de Poisson

    mod <- glmmTMB(outcome ~ temps*treat + (1|famille/id), data = famille, family = poisson)

    # Récupération de la p-value d'intérêt et agrégation au fil des itérations

    res <- res %>% bind_rows(
      broom::tidy(car::Anova(mod)) %>% filter(term == “temps:treat") %>%
        select(p.value) %>%
        mutate(nclust = nclust)
    )
  }
}
```

Exemple avec R