Consultation statistique avec le logiciel 

Les bébés sans mains de l'Ain : pour le TP MathSV ?

Pr Jean R. Lobry

Comment se forger sa propre opinion en exploitant des données publiques et une approche par simulation. L'idée serait d'utiliser ces données pour le TP sur l'échantillonnage de MathSV. On peut envisager plein de stratégies et la discussion sur le risque α est très concrète.

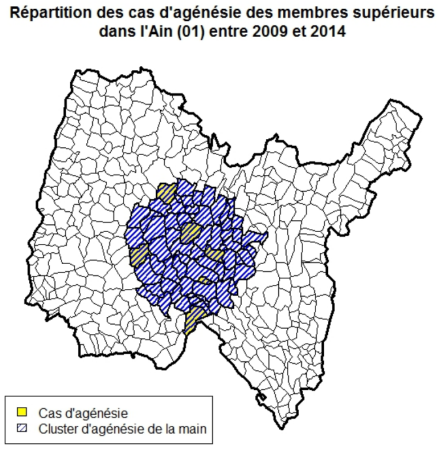
Table des matières

LES données présentées ici ont été exploitées pour construire un TP de première année de licence de science de la vie, d'une part sous la forme d'un polycopié papier traditionnel [1], et, d'autre part, sous la forme d'un document interactif [2], plus facile à mettre en œuvre dans le cadre de l'enseignement en distanciel imposé par la covid 19.

1 Introduction

1.1 Problématique

LA carte ci-dessous, extraite de la page 12 d'un rapport public [3], donne la localisation spatiale de cas d'anomalies [4] du développement des membres supérieurs chez *Homo sapiens*. C'est une étude rétrospective.



LA région d'intérêt est hachurée ci-dessus, elle comporte 67 communes réparties dans un cercle approximativement centré sur la commune de DRUILLAT et de 17.66 km de rayon. Les 7 communes pour lesquelles on a observé un cas sont en jaune, on n'en voit que 6 ici pour une raison inconnue. Un cercle de rayon 17.66 km a une surface de 979.8 km^2. Le département de l'Ain ayant une superficie de 5 762 km^2, cela représente 17 % du total. La question que l'on se pose est de savoir s'il y a une concentration anormalement élevée de cas dans cette zone : si je tire les cas au hasard, quelle est la probabilité pour qu'ils se concentrent dans une zone représentant de l'ordre de $\frac{1}{6}$ de la surface du département ?

1.2 Du choix de la latéralité du test

LES étudiants n'ont pas abordé en TD la question de la latéralité des tests statistiques. C'est l'occasion d'introduire la notion sur un cas concret. Com-

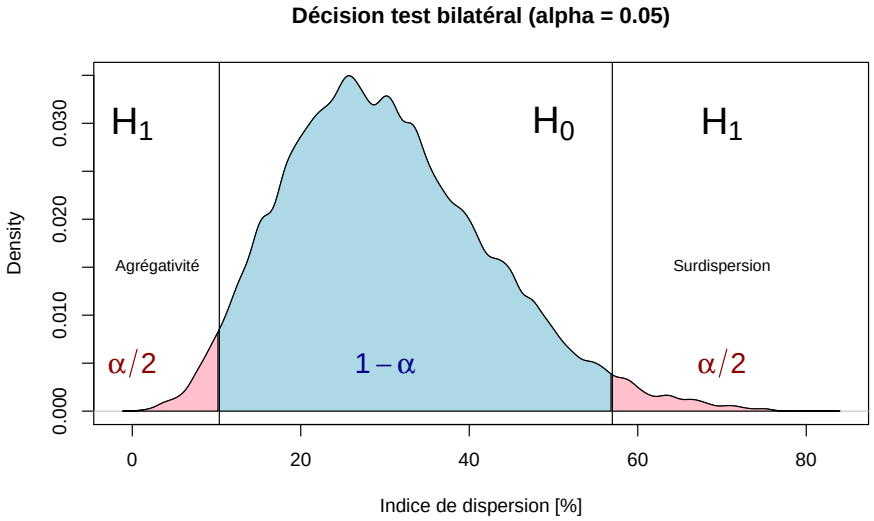
1. `http://pbil.univ-lyon1.fr/R/pdf/tpbb.pdf`
2. `http://umr5558-shiny.univ-lyon1.fr/tpbb/`
3. « Détection d'un agrégat spatio-temporel d'anomalies réductionnelles des membres chez des enfants nés dans l'Ain entre 2009 et 2014 » du Registre des Malformations en Rhône Alpes (REMERA).
4. Dans le raport de « Santé publique France » [8] ces anomalies sont désignées par l'acronyme ATMS pour « agénésies transverses des membres supérieurs ».

mençons par un test bilatéral et posons l'hypothèse nulle H_0 et son alternative H_1 :

H_0 : les cas d'anomalies apparaissent de façon indépendante de leur géolocalisation les uns des autres sur le territoire

H_1 : les cas d'anomalies n'apparaissent pas de façon indépendante de leur géolocalisation les uns des autres sur le territoire

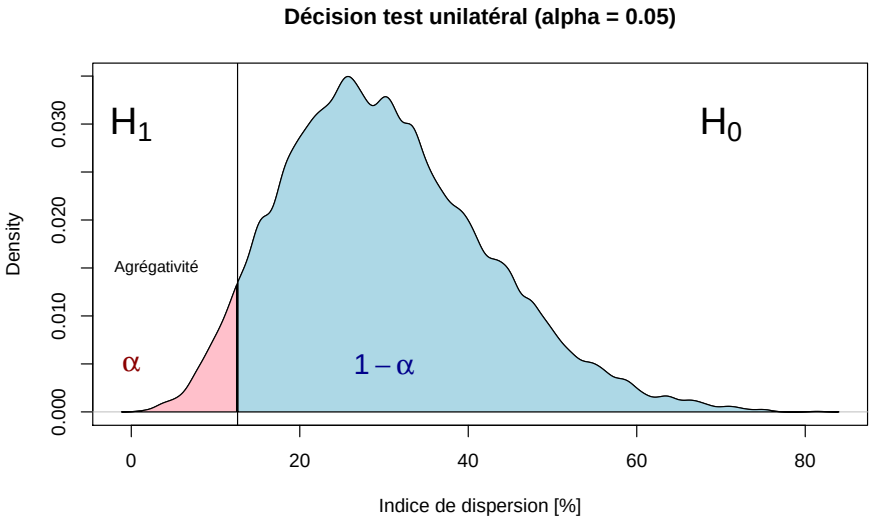
L'hypothèse alternative H_1 peut signifier plusieurs choses et en particulier soit qu'il y a une tendance à l'agrégativité des cas soit au contraire qu'il y a une sur-dispersion des cas. Supposons que l'on ait défini un indice de dispersion, exprimé en pourcentage, dont on connait la fonction de densité de probabilité sous H_0. Le test nous conduira à rejeter H_0 pour les valeurs anormalement faibles ou fortes, comme illustré ci-dessous avec $\alpha = 0.05$.



Pour les études épidémiologiques, ce qui nous intéresse ce sont plutôt les cas d'agrégativité. Par exemple, dans son étude sur la géolocalisation des cas de décès par choléra, John Snow avait trouvé une agrégation des cas dans les territoires du ressort de certaines compagnie distributrices d'eau. Mais ce n'est pas forcément le cas, si on s'intéresse à la géolocalisation d'animaux terrestres une sur-dispersion peut être l'indice intéressant d'un comportement territorial marqué. Il n'y a pas de recette de cuisine pour choisir la latéralité d'un test : tout dépend de l'*objectif* poursuivi. Ici nous voulons détecter les cas de dispersion anormalement faible, nous utilisons donc un test unilatéral :

H_0 : les cas d'anomalies apparaissent de façon indépendante de leur géolocalisation les uns des autres sur le territoire, ou sont sont éventuellement sur-dispersés.

H_1 : les cas d'anomalies sont anormalement peu dispersés sur le territoire.

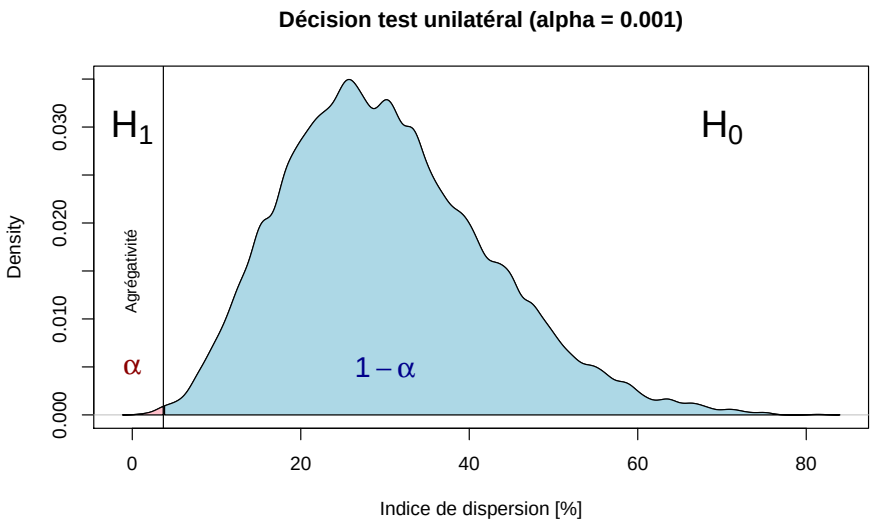


L'avantage du test unilatéral par rapport au test bilatéral c'est que le risque de première espèce, α, celui que l'on contrôle, représente le phénomène qui nous intéresse. C'est le risque de décider à tort qu'il y a une concentration anormale des cas sur le territoire.

1.3 Du choix du risque de première espèce α

Le risque de première espèce, α, c'est la probabilité de rejeter à tort l'hypothèse nulle. C'est le seul que l'on contrôle en pratique. Rejeter à tort l'hypothèse nulle c'est déclarer qu'il y a une situation anormale alors qu'il n'en est rien, et donc d'affoler les populations pour rien dans notre cas. Si on utilise le classique $\alpha = 0.05$ cela me semble bien trop élevé. Comme il y a environ 100 départements en France, si on faisait la même étude dans tous les départements, on affolerait inutilement les résidents de 5 départements. De même $\alpha = 0.01$ me semble encore un peu trop élevé, ce qui me semblerait raisonnable c'est travailler avec un risque $\alpha = 10^{-3}$. Si ce choix ne vous convient pas il suffit de recompiler cette fiche en changeant la ligne de code ci-dessous.

```
(alpha <- 1e-3)
```
```
[1] 0.001
```



Dans la fable d'Ésope « Le Garçon qui criait au loup », le protagoniste s'amuse à prétendre qu'il a vu un loup, ce qui finit par le discréditer auprès des habitants de son village. Le jour où il voit vraiment un loup, personne ne prête attention à son cri d'alarme, et il finit dans les entrailles de l'animal. Le risque de première espèce, α, c'est le risque de crier au loup, et une entité à caractère institutionnel telle que « Santé publique France » ne peut se permettre un risque trop important ici au risque de perdre sa crédibilité au près des citoyens et d'être dissoute.

Le risque de seconde espèce, β, c'est le risque de ne pas décider qu'il y a une concentration spatiale des cas alors que c'est le cas. Le problème est qu'il est impossible de minimiser simultanément les risques de première et seconde espèce. Pour s'en convaincre il suffit de considérer le cas limite « à la Ponce-Pilate » où l'on choisit $\alpha = 0$ %, dans ce cas on a un risque de seconde espèce $\beta = 100$ %. Le choix du risque de première espèce est toujours une affaire de compromis politique.

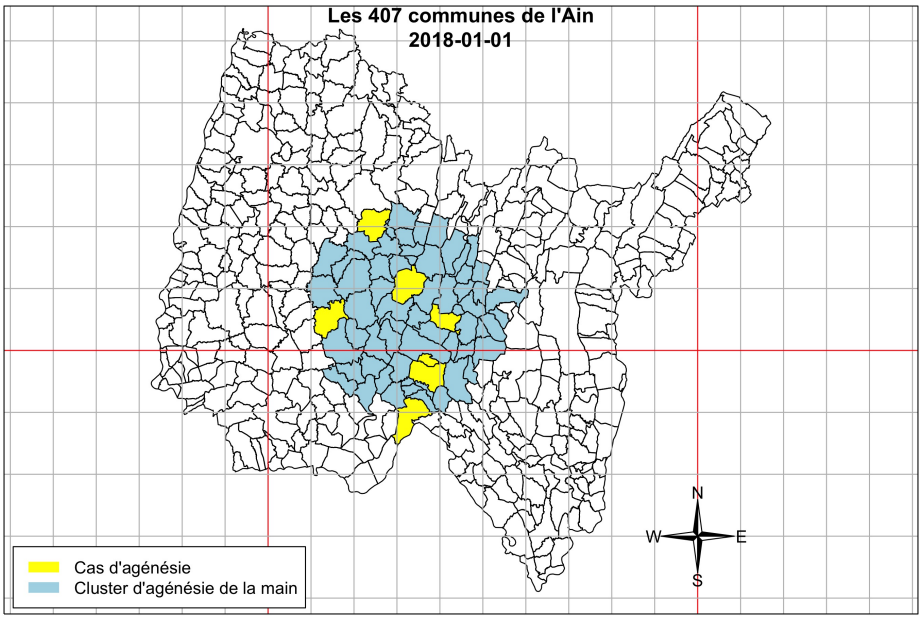
1.4 Données nécessaires

Je vais avoir besoin de récupérer les données sur la localisation spatiale des communes du département de l'Ain. Je vais avoir besoin également du nombre de naissances vivantes pendant la période considérée pour chaque commune puisque sous H_0 le nombre de cas observés est simplement proportionnel au nombre de naissances dans chaque commune.

2 Géolocalisation des communes de l'Ain

Je suis parti du fichier `communes-20180101.shp` donnant le découpage administratif communal français issu d'OpenStreetMap dont j'ai extrait les communes du département de l'Ain. Les coordonnées sont en WGS84. On les importe facilement sous R avec la fonction `readShapeSpatial()` du paquet `maptools` [1]. La fonction standard `identify()` permet très facilement de re-

pérer les communes d'intérêt en cliquant, ce qui m'a permis de rajouter des couleurs (*cf.* section 7 page 20 pour plus de détails sur l'identification des communes).



L'aspect de la carte est le même que celui du rapport, avec un rapport de 0.7 environ pour la longueur d'un méridien pour un parallèle à 45 ° de latitude. Dans la suite je résumerai les entités surfaciques des communes par les simples coordonnées du centre géométrique d'icelles, le calcul est fait par la fonction `coordinates()` du paquet `sp` [6, 2]. Voici un exemple d'utilisation.

```
library(seqinr) # pour col2alpha()
xy <- coordinates(ain)
par(mar = c(0, 0, 0, 0) + 0.1)
plot(ain, border = "transparent")
points(xy, pch = 19, cex = 0.75)
text(xy, labels = ain$nom, pos = 4, cex = 0.5, col = col2alpha("grey", 0.7))
```

3 Nombre de naissances des communes entre 2009 et 2014

On trouve facilement sur le site de l'INSEE [5] des séries longues sur le nombre de naissances vivantes domiciliées par commune. Je suis parti d'un fichier [6] dont j'ai extrait les communes du département de l'Ain.

```
naiss <- read.table("http://pbil.univ-lyon1.fr/R/donnees/qrbb/naiss.csv",
                    header = TRUE, sep = "\t", quote = "", dec = ".",
                    stringsAsFactors = FALSE)
names(naiss)
```

```
[1] "CODGEO"  "LIBGEO"  "NAISD08" "NAISD09" "NAISD10" "NAISD11" "NAISD12" "NAISD13"
[9] "NAISD14" "NAISD15" "NAISD16" "NAISD17"
```

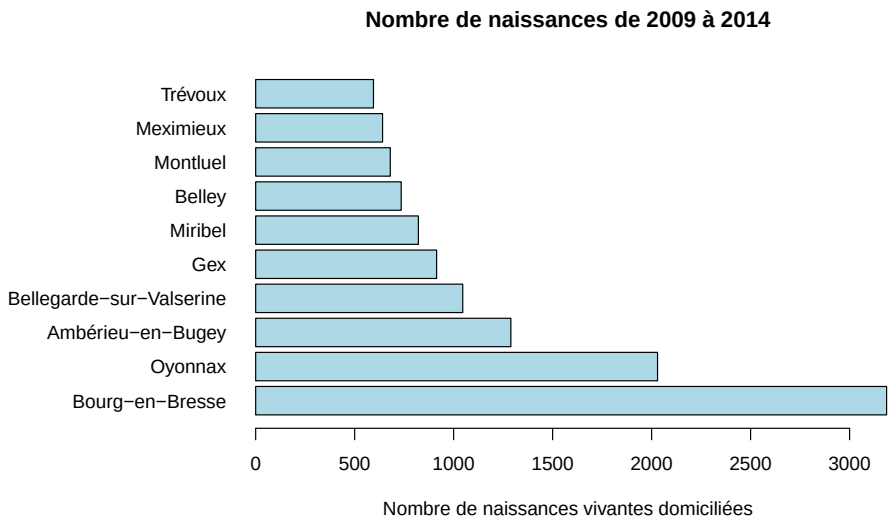
```
target <- c("NAISD09", "NAISD10", "NAISD11", "NAISD12", "NAISD13", "NAISD14")
itarget <- which(colnames(naiss) %in% target)
naiss$naiss <- rowSums(naiss[, itarget])
sum(naiss$naiss)
```

```
[1] 43596
```

Il y a de l'ordre de 7 000 naissances par an dans le département de l'Ain, le nombre total trouvé ici pour 6 années semble correct. Voyons quelles sont les communes les plus prolifiques pour vérifier qu'il n'y a pas d'erreur grossière ici.

```
x <- naiss$naiss
names(x) <- naiss$LIBGEO
par(mar = c(5, 11, 4, 1) + 0.1)
barplot(rev(sort(x))[1:10], horiz = TRUE, las = 1, col = "lightblue",
        main = "Nombre de naissances de 2009 à 2014",
        xlab = "Nombre de naissances vivantes domiciliées")
```

5. https://www.insee.fr/
6. base_naissances_2017.xls



Je ne vois rien de foncièrement aberrant dans ces données, la préfecture du département représente le plus grand nombre de naissances, ce qui semble assez logique.

4 Jointure des deux tables

4.1 Vérification de la cohérence des clefs et jointure

```
load(url("http://pbil.univ-lyon1.fr/R/donnees/qrbb/AinCol.Rda"))
# Est-ce que code_insee est bien une clef d'identification
ain$insee <- as.integer(as.character(ain$insee))
any(duplicated(ain$insee)) # FALSE, OK
```

```
[1] FALSE
```

```
# Est-ce que CODGEO est bien une clef d'identification
any(duplicated(naiss$CODGEO)) # FALSE, OK
```

```
[1] FALSE
```

```
# Puis-je joindre les tables sans perte d'information ?
all(ain$insee %in% naiss$CODGEO) # TRUE, OK
```

```
[1] TRUE
```

```
# Je fais la jointure
dta <- merge(ain, naiss[ , c("CODGEO", "naiss")],
             by.x = "insee",
             by.y = "CODGEO")
save(dta, file = "dta.Rda")
```

4.2 Exploitation de la table

```
load(url("http://pbil.univ-lyon1.fr/R/donnees/qrbb/dta.Rda"))
dta.df <- as.data.frame(dta)
with(dta.df, sum(naiss[col == "yellow"]))
```

```
[1] 5413
```

```
p <- 1.7/10000
signif(p*sum(dta.df$naiss), 3)
```

```
[1] 7.41
```

```
dta.df$theo <- round(p*dta.df$naiss, 2)
dta.df[dta.df$col == "yellow", c("insee", "nom", "naiss", "theo")]
```

```
    insee                    nom naiss theo
135  1149                Douvres    60 0.01
3    1004      Ambérieu-en-Bugey  1289 0.22
180  1202                Lagnieu   533 0.09
178  1199              Jujurieux   150 0.03
337  1381 Saint-Nizier-le-Désert    79 0.01
331  1374   Saint-Martin-du-Mont   115 0.02
49   1053        Bourg-en-Bresse  3187 0.54
```

Au niveau du département, j'ai un total de 43596 naissances vivantes pour la période considérée dans mon jeu de données. Sachant qu'il y a de l'ordre de $7\ 10^3$ naissances par an dans l'Ain, c'est parfaitement cohérent avec une période de 6 années.

I TABLEAU 1 I

Calcul du rapport d'incidence (SIR) des cas d'agénésie de membres supérieurs dans les communes concernées entre 2009 et 2014

Lieu de domicile	Nombre de cas observés d'agénésie (REMERA)	Nombre de cas attendus	SIR [IC95] (approximation de Byar [3])
Commune 1	1	0,54	1,85 [0,02-10,27]
Commune 2	1	0,04	24,31[0,32-135,24]
Commune 3	1	0,03	31,29 [0,41-174,08]
Commune 4	1	0,03	39,22 [0,51-218,19]
Commune 5	1	0,09	11,04 [0,14-61,40]
Commune 6	1	0,22	4,56 [0,06-25,40]
Commune 7	1	0,03	37,71 [0,49-209,80]
Ain	**7**	**7,41**	**0,94 [0,38-1,95]**

Si on multiplie le nombre total de naissances dans mon jeu de données avec la probabilité d'occurrence, $p = 1.7\ 10^{-4}$, on trouve exactement la même valeur, 7.41, que celle donnée dans la table de « Santé publique France » pour le nombre de cas attendus sur la période, et ce avec trois chiffres significatifs. Ceci me conforte dans l'idée que pour la période entre 2009 et 2014 c'est bien de 2009 à 2014, *inclus*.

Au niveau du département, il n'y a donc rien d'extraordinaire que d'observer 7 cas d'agénésie sur une période de 6 ans puisque l'on s'attendrait à en observer 7.41 en moyenne. On peut être un peu plus précis ici en modélisant le phénomène à l'aide d'un schéma de Bernoulli : on fait $n = 43\ 596$ tirages avec remise dans une urne contenant des bébés et pour chaque tirage i on définit la variable aléatoire X_i indicatrice des cas qui nous intéressent :

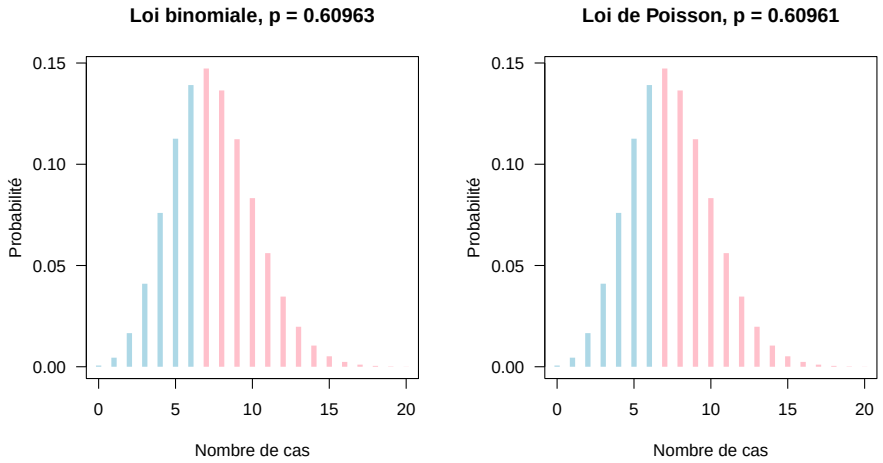
$$\begin{cases} X_i &= 1 \text{ si le bébé présente une agénésie} \\ X_i &= 0 \text{ sinon} \end{cases}$$

On note $p = P(X_i = 1) = 1.7\ 10^{-4}$ la probabilité d'observer un cas et on définit la variable aléatoire Y :

$$Y = \sum_{i=1}^{n} X_i$$

Qui représente le nombre total de cas observés sur l'ensemble du département pendant la période considérée. En tant que somme de variables aléatoires

de Bernoulli supposées toutes indépendantes deux à deux, Y suit une loi binomiale de paramètres n et p, que l'on peut approximer par une loi de Poisson de paramètre $\lambda = np$. Ceci nous permet de calculer les probabilités d'observer un nombre de cas donné :



Il y a donc 61 % de chances d'observer 7 cas ou plus à l'échelle du département, il n'y a donc vraiment rien d'exceptionnel ici. Avec un risque de première espèce $\alpha = 10^{-3}$ il faudrait observer **17 cas** pour décider que c'est un nombre anormalement élevé :

```
 1 - sum(dbinom(0:16, n, p))
[1] 0.001733694
 1 - sum(dbinom(0:17, n, p))
[1] 0.0006911667
```

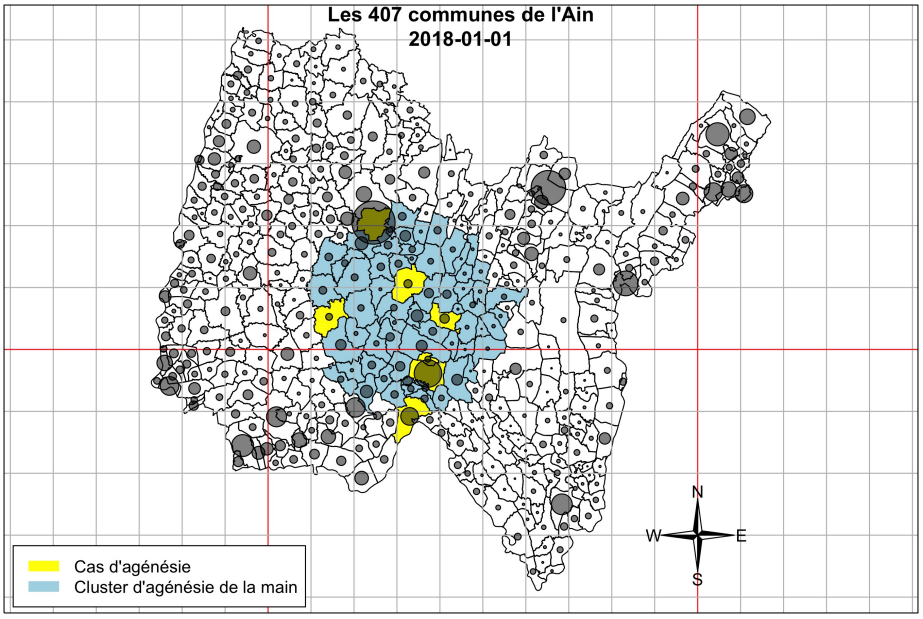
La conclusion [8] de « Santé publique France » d'octobre 2018 « [c]es rapports d'incidence ne sont pas statistiquement significatifs, dans les 7 communes concernées par des cas d'ATMS et au niveau départemental » est parfaitement justifiée à l'échelle du département. Mais ceci ne nous dit rien sur la probabilité pour que ces cas soient concentrés sur une zone représentant $\frac{1}{6}$ de la surface du département.

4.3 Calcul naïf de la probabilité de concentration des cas

Si on suppose une répartition uniforme des naissances sur le territoire, la probabilité qu'ils soient tous dans cette zone est de $\left(\frac{1}{6}\right)^7 = 3.6\ 10^{-6}$. On peut être un peu plus précis ici puisque dans le jeu de données on dispose de la colonne `surf_ha` qui donne la surface des communes.

```
 sz <- with(dta.df, sum(surf_ha[col != "transparent"])) # surface de la zone
 (ps <- sz/sum(dta.df$surf_ha))
[1] 0.1744592
 1/ps
[1] 5.731998
 ps^7
[1] 4.918782e-06
```

Cinq chances sur un million, ce n'est vraiment pas beaucoup. Bien entendu l'hypothèse de répartition uniforme des naissances sur le territoire mérite d'être examinée de plus près :



La répartition des naissances est loin d'être uniforme sur le territoire, il y a une concentration sur la préfecture et un effet lisière avec les bassins de Genève, Lyon et de la « plastic vallée ». Si on raffine un peu en tenant compte cette fois du nombre de naissances :

```
nz <- with(dta.df, sum(naiss[col != "transparent"])) # naissances de la zone
(pz <- nz/sum(dta.df$naiss))
```

```
[1] 0.2314433
```

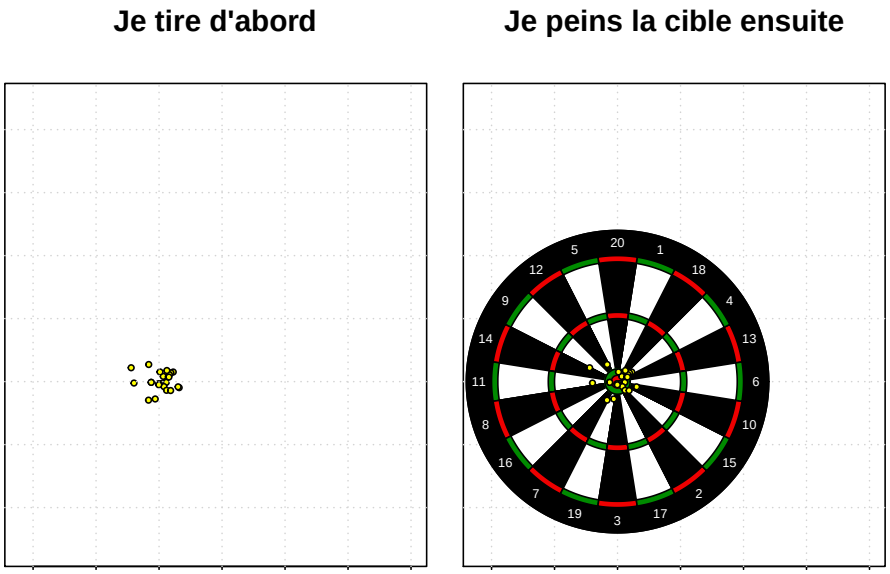
```
1/pz
```

```
[1] 4.320714
```

```
pz^7
```

```
[1] 3.557228e-05
```

Comme la zone contient la préfecture de Bourg-en-Bresse où il y a beaucoup de naissances, la probabilité d'avoir une naissance domiciliée dans la zone est supérieure à $\frac{1}{6}$, plutôt de l'ordre de $\frac{1}{4}$, et la probabilité pour que tous les cas y soient est de $3.6\ 10^{-5}$. Voila qui pourrait sembler fort inquiétant : en tirant les cas au hasard il y a un probabilité très faible pour qu'ils se concentrent dans la zone. Sauf que cela ne veut strictement rien dire parce que la zone a été définie *a posteriori* en baladant un cercle sur le département et en cherchant à maximiser le nombre de cas à l'intérieur du cercle. C'est le syndrome dit du « tireur d'élite texan » qui consiste à tirer sur un mur et à peindre la cible ensuite :

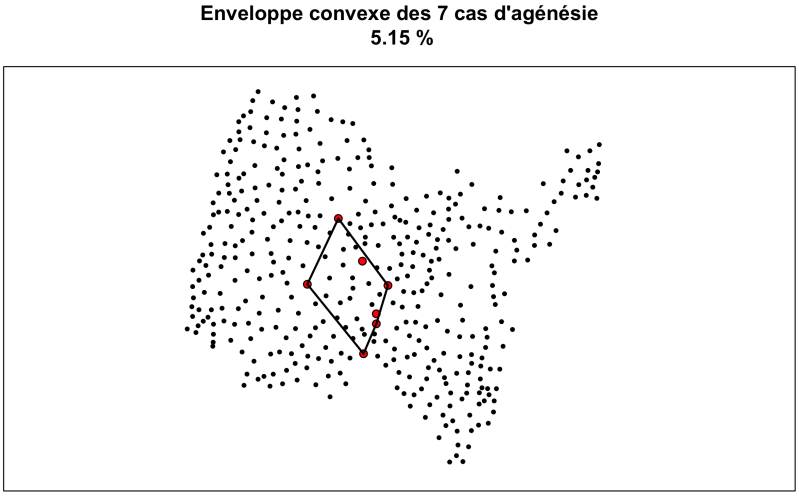


ON pourrait être fort impressionné par les performances de notre tireur d'élite texan si on ne savait pas que la cible a été dessinée *a posteriori*. Si on tient à apprécier les performances de notre tireur d'élite texan il faudrait répéter l'expérience avec de nombreux autres tireurs dans les mêmes conditions. C'est l'idée de l'approche par simulation que l'on utilisera ici.

5 Approche par simulation

5.1 Choix d'une statistique

ON veut savoir s'il y a une concentration anormalement élevée de cas sur une zone géographique. Il nous faut déjà définir une statistique pour caractériser le degré de concentration géographique des cas. Il y a plusieurs façons de le faire, j'ai retenu celle de la surface de l'enveloppe convexe des points parce qu'elle est assez intuitive et simple à comprendre. Voici l'enveloppe convexe des communes pour lesquelles on a observé les 7 cas d'agénésie :



L'enveloppe convexe est simplement le polygone que l'on obtient avec un élastique qui enserre tous les points. On peut calculer sa surface et l'exprimer, pour faciliter l'interprétation, en pourcentage de la surface de l'enveloppe convexe de toutes les communes du département. On a ainsi une statistique qui varie entre 0 et 100 %. Une valeur faible signifie que l'on a une concentration spatiale des cas, par exemple dans le cas extrême 0 % si tous les cas sont observés sur une seule commune. Une valeur élevée signifie que l'on a une dispersion spatiale des cas, par exemple 100 % si les cas sont dispersés aux « quatre coins [7] » du département.

La statistique pour les 7 cas d'agénésie vaut donc 5.15 %. Que penser de cette valeur ? Dans l'absolu, rien, comme pour le tireur d'élite texan on a besoin de point de comparaison pour pouvoir se prononcer. On aimerait savoir si elle est anormalement faible par rapport à ce qui serait attendu si on tirait les cas au hasard.

Note pour un futur TP : j'ai utilisé ici la fonction `ahull()` du paquet `alphahull` [5] parce qu'il y avait la fonction associée `areaahulleval()` pour calculer la surface de l'enveloppe convexe. Pour le TP il vaudrait mieux utiliser la fonction `chull()` qui est livrée avec R [7] en standard et implémente l'algorithme d'Eddy [4, 3]. Reste à trouver la fonction standard pour calculer la surface d'un polygone convexe.

```
areapoly <- function(xy){
  xy <- rbind(xy, xy[1, ])
  n <- nrow(xy) ; x <- xy[ , 1] ; y <- xy[ , 2]
  return(-0.5*sum(x[-n]*y[-1] - x[-1]*y[-n]))
}
load(url("http://pbil.univ-lyon1.fr/R/donnees/qrbb/dta.Rda"))
CRS1 <- CRS("+proj=longlat +ellps=WGS84 +datum=WGS84 +no_defs")
proj4string(dta) <- CRS1
xy <- coordinates(dta)
plot(xy, pch = 19, cex = 0.5)
cas <- which(dta$col == "yellow")
```

7. L'enveloppe convexe de l'ensemble des communes du département de l'Ain comporte 14 points, on ne pourra donc pas couvrir 100 % de sa surface avec seulement 7 points.

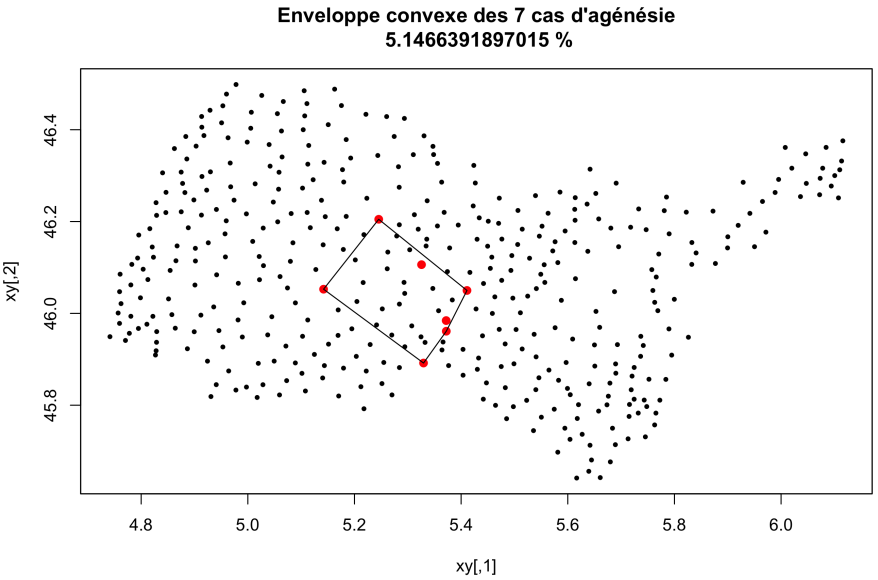
```
points(xy[cas, ], pch = 19, col = "red")
ihull <- chull(xy[cas, ])
polygon(xy[cas[ihull], ])
tots <- areapoly(xy[chull(xy),])
print(tots)
```

```
[1] 0.8263557
```

```
print(areapoly(xy[cas[ihull],]))
```

```
[1] 0.04252955
```

```
pctot <- 100*areapoly(xy[cas[ihull],])/tots
main <- paste("Enveloppe convexe des 7 cas d'agénésie\n", signif(pctot, 15), "%")
title(main = main)
box()
```



Je retrouve les mêmes valeurs, mais le calcul n'est pas tout à fait correct puisque l'on ne tient pas compte de la projection avec `coordinates(xy)`. Cela ne devrait pas changer grand chose puisque l'on travaille en surfaces relatives, mais autant être précis. Comme on veut éviter d'utiliser le packet `sp` en TP parce qu'il n'est pas de base, il faut prétraiter les données pour les étudiants. Quelle est la longitude du milieu de la carte ?

```
(My <- 0.5*(max(xy[ , 2] + min(xy[ , 2]))))
```

```
[1] 46.0699
```

```
# Voir doc de sp:::mapasp
(sc <- cos(My*pi/180))
```

```
[1] 0.6937803
```

À la latitude du département un arc d'angle x de longitude vaut environ 0.694 arc d'angle x de latitude. Pour faciliter la vie des étudiants on met à l'échelle :

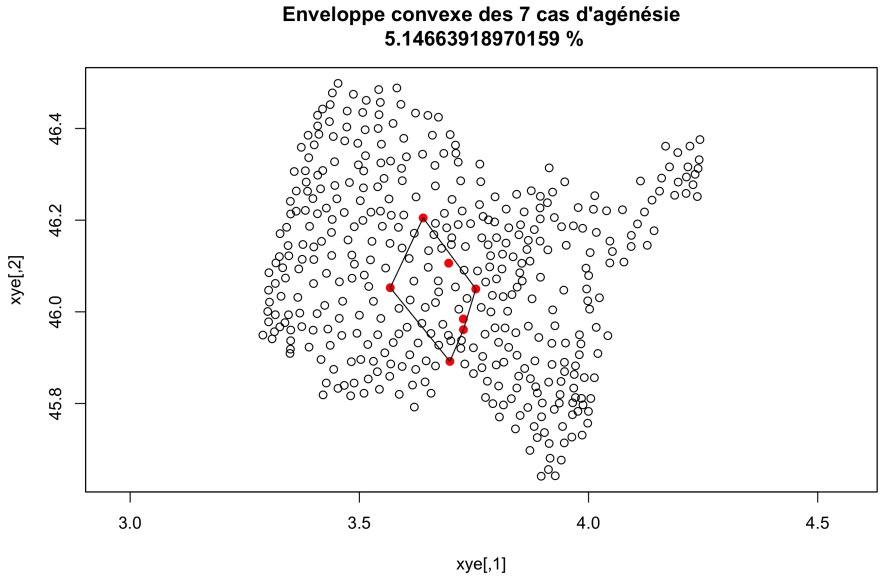
```
xye <- xy
xye[ , 1] <- sc*xye[ , 1]
plot(xye, asp = 1)
cas <- which(dta$col == "yellow")
points(xye[cas, ], pch = 19, col = "red")
ihull <- chull(xye[cas, ])
polygon(xye[cas[ihull], ])
tots <- areapoly(xye[chull(xye),])
print(tots)
```

```
[1] 0.5733093
```

```
print(areapoly(xye[cas[ihull],]))
```

```
[1] 0.02950616
```

```
pctot <- 100*areapoly(xye[cas[ihull],])/tots
main <- paste("Enveloppe convexe des 7 cas d'agénésie\n", signif(pctot, 15), "%")
title(main = main)
box()
```



C'est effectivement négligeable. Pour les étudiants on gardera les données de `xye` pour faire les cartes avec `asp = 1`.

5.2 Tirage au hasard

On définit une petite fonction `simu()` pour tirer au hasard une commune (son rang dans le tableau pour être précis) en pondérant par le nombre de naissance de chaque commune. Pour comparer des choses comparables, on fait ces tirages conditionnellement au fait que l'on sait qu'il y a eu 7 cas observés sur le département [8].

```
simu <- function() sample(x = 1:nrow(dta), size = 7, replace = TRUE, prob = dta$naiss)
simu()
```

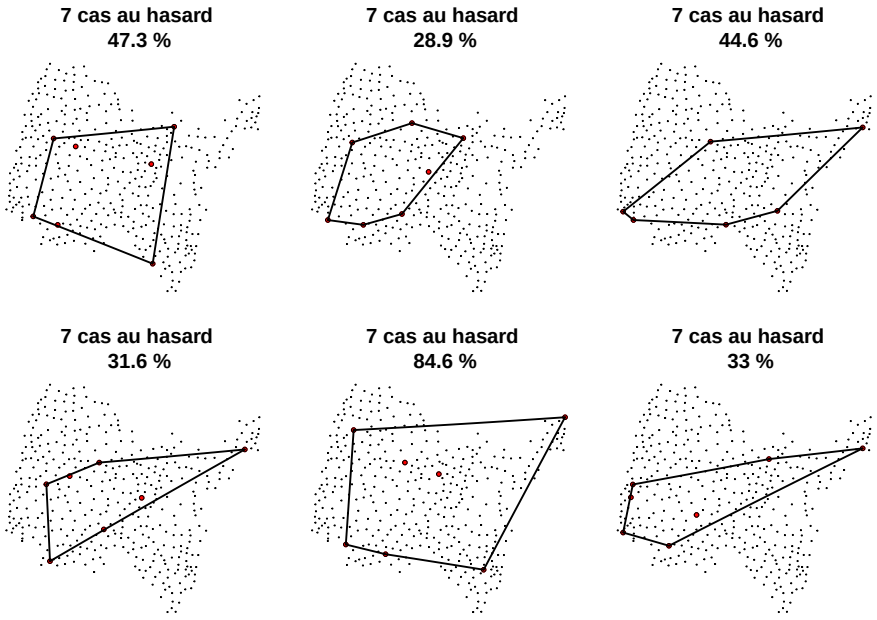
```
[1] 323  33 149 330  81 182 118
```

Ce qui est important ici c'est que les tirages au hasard se font indépendamment de la localisation géographique des communes : la probabilité d'avoir un cas dans une commune n'est pas modifiée par la présence d'un cas dans une commune limitrophe. Le seul facteur qui joue c'est le nombre de naissances dans la commune.

8. On pourrait facilement faire une étude plus générale en tirant le nombre de cas observés dans une loi de Poisson de paramètre $\lambda = np$ où n est le nombre total de naissances (43596) et p la probabilité d'une agénésie ($p = 1.7\ 10^{-4}$). Mais attention, la statistique est dépendante du nombre de cas, par exemple avec un seul cas on a forcément 0 % pour la proportion de la surface totale. Donc autant se restreindre directement au cas qui nous intéresse.

On va maintenant faire un grand nombre de simulations. Commençons par en faire 6 comme dans la figure ci-après pour expliquer le principe de la démarche :

```
manip <- function(plot = FALSE){
  if(plot) plot(dta, border = "transparent")
  if(plot) points(xy, pch = 19, cex = 0.25)
  cas <- unique(simu())
  if(plot) points(xy[cas, ], pch = 19, col = "red")
  hull <- try(ahull(xy[cas, ], alpha = alphahull), silent = TRUE)
  if(inherits(hull, "try-error")) return(NA)
  if(plot) plot(hull, add = TRUE)
  pctot <- 100*areaahulleval(hull)/tots
  if(plot) main <- paste("7 cas au hasard\n", signif(pctot, 3), "%")
  if(plot) title(main = main, cex.main = 2)
  return(pctot)
}
set.seed(1)
par(mfrow = c(2, 3), mar = c(1, 0, 3, 0) + 0.5)
res <- replicate(6, manip(plot = TRUE))
```



On voit que sur ces 6 simulations la valeur de notre statistique gambade entre 14.7 % et 40.3 %, la valeur observée de 5.15 % avec les cas réels semble donc plutôt petite. Mais pour affiner les choses il va falloir faire plus de simulations. On évitera d'afficher le graphique à chaque fois.

```
res <- replicate(10^4, manip())
save(res, file = "res.Rda")
```

```
load(url("http://pbil.univ-lyon1.fr/R/donnees/qrbb/res.Rda"))
res <- res[!is.na(res)]
range(res)
```

```
[1]  1.418185 81.407628
```

La première constatation c'est que dans nos 10 000 simulations, la valeur de notre statistique gambade entre 1.42 et 81.4 %, il est donc parfaitement

possible d'obtenir une valeur inférieure ou égale à 5.15 % *par le pur fruit du hasard*. Maintenant, est-ce que cela arrive souvent que d'avoir une valeur inférieure ou égale à 5.15 % par hasard ?

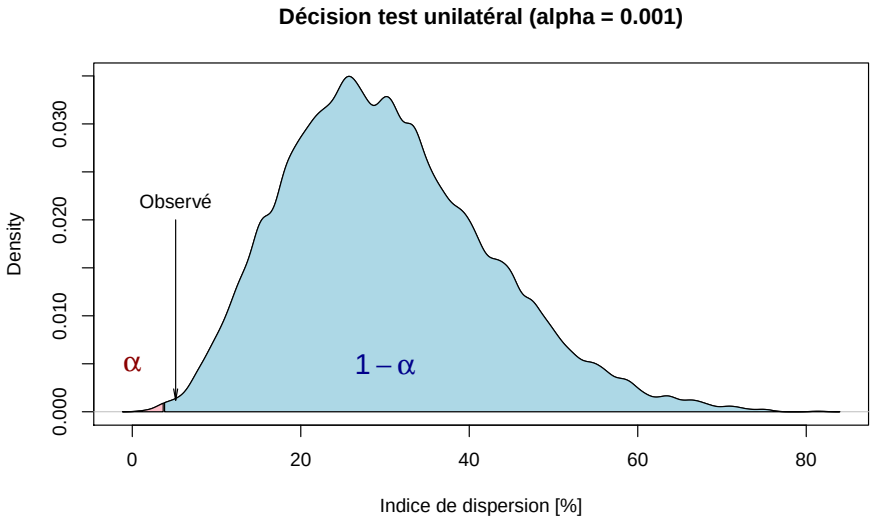
```
res[res <= 5.15]
```

```
 [1] 5.051283 4.190001 3.064863 3.688372 4.439597 4.074967 3.159260 3.614615 4.811881
[10] 4.609469 3.642276 3.867835 3.888398 4.940314 2.512652 1.418185 4.995411 3.131053
[19] 4.045080 4.277785 4.384033 1.548109 4.057104 3.357170 4.981389 3.236220 5.146552
```

```
sum(res <= 5.15)
```

```
[1] 27
```

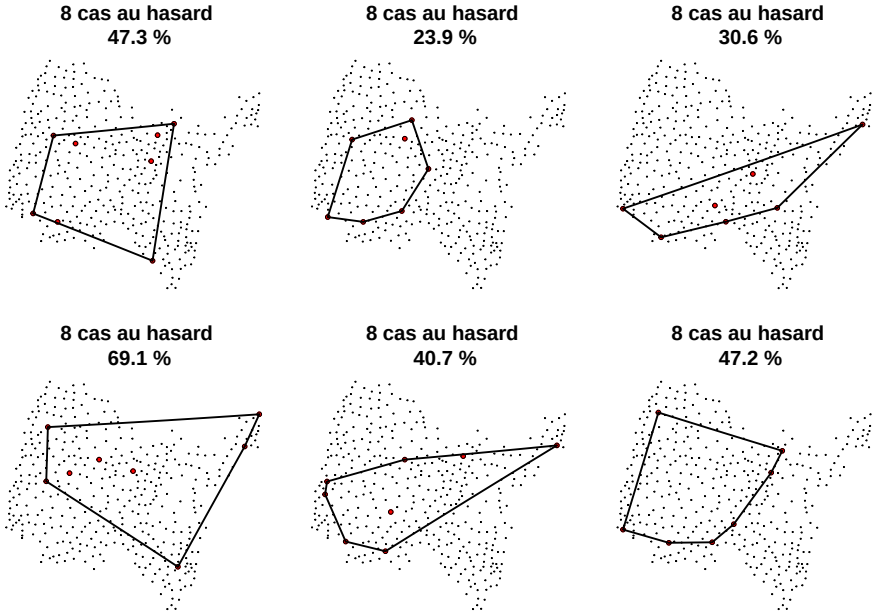
Pour 27 simulations on a obtenu par hasard une valeur de la statistique inférieure ou égale à 5.15 %, comme on a fait 10 000 simulations en tout, la probabilité d'une telle situation est donc de l'ordre de $2.7\ 10^{-3}$. Je vais donc maintenant pouvoir prendre une décision. Je me suis fixé un risque de première espèce $\alpha = 10^{-3}$, je suis au dessus de cette valeur critique, donc je suis dans l'incapacité de rejeter l'hypothèse nulle, je l'accepte donc, faute de mieux, avec un risque de seconde espèce β inconnu. Je ne suis pas arrivé à mettre en évidence une agrégativité spatiale des cas et je ne contrôle pas le risque associé à une telle décision, c'est la situation la plus inconfortable, comme d'habitude les tests statistiques ne sont probants qu'au rejet.



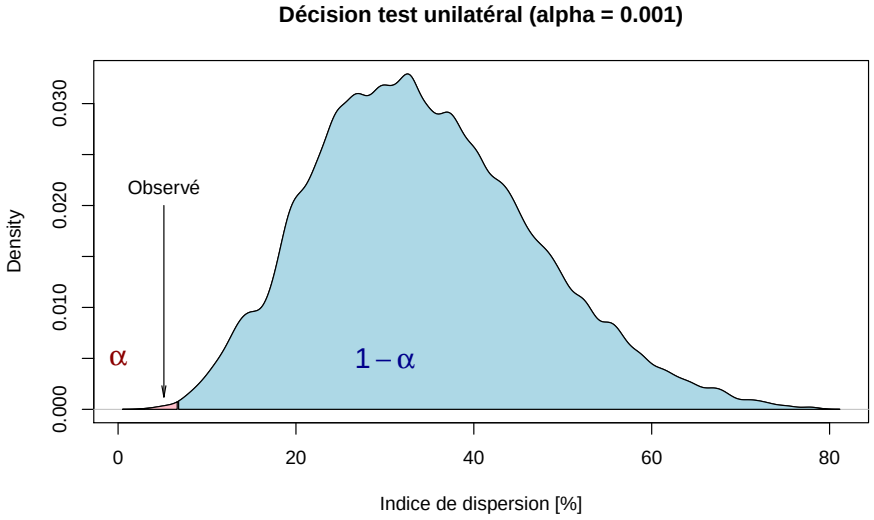
5.3 Breaking news

Dans un article [9] du journal « Le Monde » du 29 octobre 2018, un huitième cas est signalé dans la zone d'intérêt. Sous réserve que ce cas fasse bien parti des critères d'inclusion de « Santé publique France », à savoir « [t]out cas d'agénésie transverse isolée des membres supérieurs sans anomalie chromosomique connue, né et domicilié entre 2000 et 2014 dans le département de l'Ain », que devient notre test ? Comme la commune de ce nouveau cas n'est pas documentée, on va supposer pour être conservatif qu'elle est dans l'enveloppe convexe des 7 cas précédents. On relance les simulations avec 8 cas observés.

9. https://www.lemonde.fr/sante/article/2018/10/30/bebes-sans-bras-de-l-ain-un-huitieme-cas-identifie_5376404_1651302.html



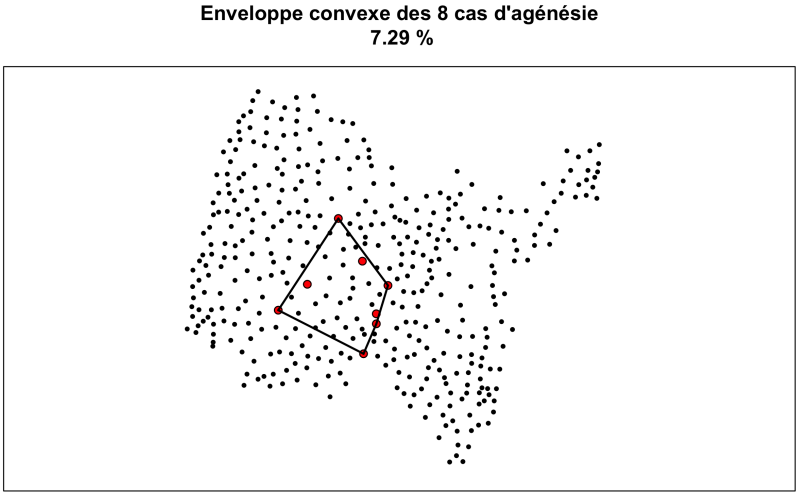

```
res <- replicate(10^4, manip8())
save(res, file = "res8.Rda")
```



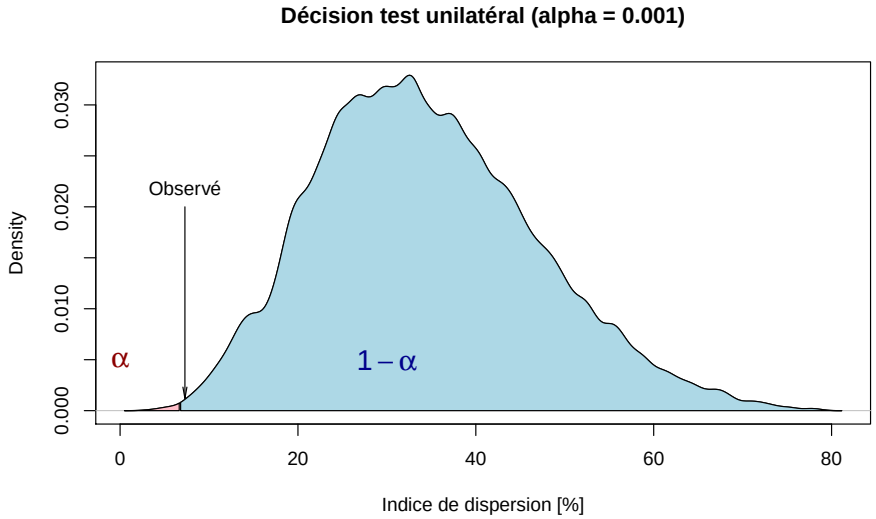
Cette fois on rejette l'hypothèse nulle H_0 : avec un risque de première espèce $\alpha = 10^{-3}$ les données expérimentales sont en contradiction avec l'hypothèse d'une répartition spatiale aléatoire des cas observés. Le seul bémol ici est que l'on ne connait pas la nouvelle commune, la statistique de la surface convexe des cas observée est peut être supérieure à 5.15 %.

D'après le journal « Le parisien [10] », le huitème cas est dans la commune de Villars-les-Dombes. On peut recalculer la satistique des cas observés :

10. http://www.leparisien.fr/societe/sante/bebes-nes-sans-bras-a-villars-les-dombes-on-s-interroge-31-10-2018.php



La commune de Villars-les-Dombes n'était pas dans l'enveloppe convexe des 7 cas précédent, ce qui augmente donc la valeur de la statisqtique des cas observés, on rebascule du coté de H_0 :



6 Discussion

Je pense qu'avec ce jeu de données on pourrait renouveler le TT-arbre pour leur faire pratiquer l'échantillonnage. L'avantage ici c'est que l'on a des données réelles, avec une problématique concrète facile à comprendre : est-ce qu'il y a oui ou non une agrégativité spatiale des cas d'agénésie dans l'Ain ? Je note quelques idées en vrac.

— Problématique. On leur fait faire le calcul naïf de la proba d'avoir 7 points concentrés sur $\frac{1}{6}$ de la surface du département.
— On leur donne la carte cliquable des centre géométriques des communes. Il en tirent 7 au hasard, ça leur renvoie la statistique plus le visuel de l'enveloppe convexe. Re-calcul naïf de la proba pour les points d'être dans l'enveloppe convexe. Rhazut ! Ça va être de l'ordre de 10^{-3}. Problème. Simuler un tir au tableau puis dessiner une cible après pour expliquer le syndrome du tireur d'élite texan. Rétrospective *vs* prospective. Intérêt méthode par simulation.
— Intérêt d'un test unilatéral ici.
— Il font un échantillon de 30. Calculent la moyenne. Comparent avec les autres sous-groupes. Qui a raison ? Critique de la méthode : fastidieux, pas aléatoire.
— On tire au hasard dans une loi uniforme les coordonnées GPS. Critique de la méthode : biais en faveur des communes de grande surface.
— On tire au hasard dans une loi uniforme les noms des communes. Critique de la méthode : biais en faveur des communes de faible natalité.
— On tire au hasard en pondérant par la natalité. Analyse de la fonction de densité empirique de la statistique sous H_0
— Mais c'est quoi H_0 au fait ?
— Leur laisser faire le test avec $\alpha = 0.05$ puis leur demander d'expliciter les risques de première et seconde espèce du point de vue des risques sanitaires. $\alpha = 0.05$ trop grand, oui mais jusqu'où descendre ? Avec une *p-value* aussi borderline, ça peut donner des discussions intéressantes.

7 Annexe

1° Pour Bourg-en-Bresse j'ai 3187 naissances ce qui me donne en multipliant par 1.7 10^{-4} un nombre cas théoriques attendu de 0.54 sur la période. C'est exactement la même valeur, avec deux chiffres significatifs, que celle donnée pour « Commune 1 » dans la table de « Santé publique France ». Bourg-en-Bresse est la commune qui porte le plus de naissances, suivie par Oyonnax avec 2030 naissance donnant 0.35 cas attendus. J'ai donc identifié de façon certaine « Commune 1 » à Bourg-en-Bresse. Bourg-en-Bresse est bien visible en jaune sur la carte du REMERA.

2° Pour Ambérieu-en-Bugey j'ai 1289 naissances, donnant 0.22 cas attendus, soit exactement la même valeur, avec deux chiffres significatifs, que dans la table pour « Commune 6 ». Ambérieu-en-Bugey est encadrée par Oyonnax avec 0.35 cas attendus et Bellegarde-sur-Valserine avec 0.18 cas attendus. J'ai donc identifié de façon certaine « Commune 6 » à Ambérieu-en-Bugey. Le problème est qu'Ambérieu-en-Bugey n'est pas en jaune sur la carte du REMERA. Je ne sais pas pourquoi, mais d'un autre coté il n'y a que 6 communes en jaune sur la carte du REMERA.

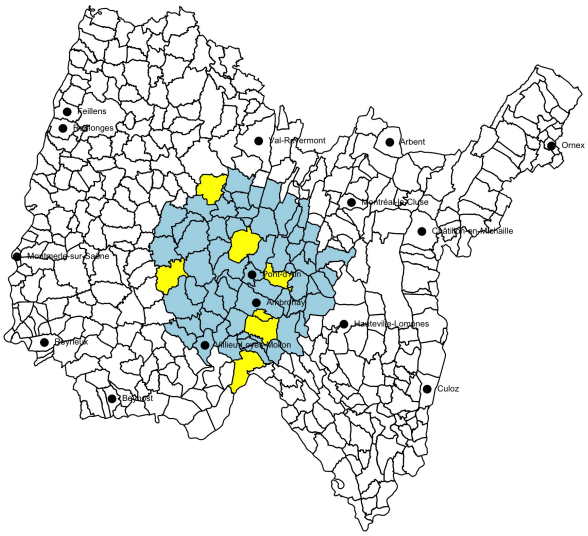
3° Pour Lagnieu j'ai 533 naissances, donnant 0.09 cas attendus, soit exactement la même valeur, avec un chiffres significatifs, que dans la table pour « Commune 5 ». Lagnieu est encadrée par Saint-Genis-Pouilly (0.10), Ferney-Voltaire (0.09) et Divonne-les-Bains (0.07). Elles sont toutes dans le pays de Gex, à l'extrême est du département, donc complètement hors-

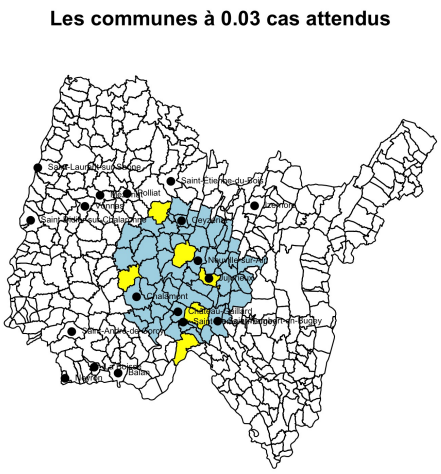
zone. J'ai donc identifié de façon certaine « Commune 5 » à Lagnieu. Lagnieu est bien visible en jaune au sud de la zone de la carte du REMERA.

4° Pour Jujurieux j'ai 150 naissances, donnant 0.03 cas attendus (0.0255 pour être précis). Je serai tenté de l'assigner à la plus grosse valeur résiduelle du tableau, 0.04 pour « Commune 2 », mais cela fait un écart de l'ordre de 60 naissances quand même. J'ai 15 communes qui sont à 0.04 cas attendus dans mon jeu de données, dont 3 sur zone : Pont d'Ain, Ambronay et Villieu-Loyes-Mollon, mais aucune ne pourrait correspondre à la carte du REMERA. Pour les communes à 0.03 cas attendus j'en ai 7 sur zone, mais il n'y a que Jujurieux qui colle avec la carte du REMERA. Je fais donc une identification putative de « Commune 2 » à Jujurieux.

5° Pour les communes restantes je n'ai pas assez de précision dans la table pour pouvoir faire une identification certaine, mais elles sont bien visibles sur la carte du REMERA.

Les communes à 0.04 cas attendus





JE ne suis donc pas arrivé à identifier de façon certaine toutes les communes du tableau de « Santé publique France ». Ce n'est pas vraiment gênant puisque l'on a une approche par simulation essentiellement spatiale.

Références

[1] Roger Bivand and Nicholas Lewin-Koh. *maptools : Tools for Handling Spatial Objects*, 2018. R package version 0.9-4.

[2] Roger S. Bivand, Edzer Pebesma, and Virgilio Gomez-Rubio. *Applied spatial data analysis with R, Second edition*. Springer, NY, 2013.

[3] W.F. Eddy. Algorithm 523 : Convex, a new convex hull algorithm for planar sets [z]. *ACM Transactions on Mathematical Software*, 3 :411–412, 1977.

[4] W.F. Eddy. A new convex hull algorithm for planar sets. *ACM Transactions on Mathematical Software*, 3 :398–403, 1977.

[5] Beatriz Pateiro-Lopez, Alberto Rodriguez-Casal, and . *alphahull : Generalization of the Convex Hull of a Sample of Points in the Plane*, 2016. R package version 2.1.

[6] Edzer J. Pebesma and Roger S. Bivand. Classes and methods for spatial data in R. *R News*, 5(2) :9–13, November 2005.

[7] R Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[8] Saint-Maurice. Investigation d'une suspicion d'agrégat spatio-temporel de malformations congénitales dans le département de l'Ain. Technical report, Santé publique France, 2018.