Projet de biostatistiques - Ankylostome

Amandine LIAGRE - Florian BUCQUET - Rachid ABDELJABBAR

03-02-2024

Contents

Introduction

Contexte

A travers ce projet, nous allons utiliser les données provenant d'une enquête réalisée sur un échantillon d'individus en Egypte. Plus particulièrement, nous avons des informations concernant l'infection des individus par l'ankylostome. Il s'agit d'un parasite intestinal. En marchant pieds nus, les individus sont directement contaminés via les larves des ankylostomes vivant en terre. L'infection peut aussi se produire via une ingestion d'aliments contaminés par des larves. Les différents symptômes possibles sont des éruptions et lésions cutanées aux endroits où les larves ont pénétré la peau, de la fièvre, des douleurs épigastriques, des diarrhées, de la toux, inflammation de l'intestin ... . Dans les cas les plus graves, le malade peut être victime d'une perte de sang (les larves dans l'intestin se nourrissent de sang en étant accroché à sa paroi et il en résulte une potentielle anémie pour le malade) et d'insuffisance cardiaque. Il existe des médicaments antiparasitaires pour traiter cette infection (albendazole, mébendazole).

L'ankylostome vit particulièrement bien dans la terre (plus précisément les sols humides) et une température aux alentours des 18°C afin que les oeufs puissent éclore. Les oeufs d'ankylostomes ont l'allure suivante:

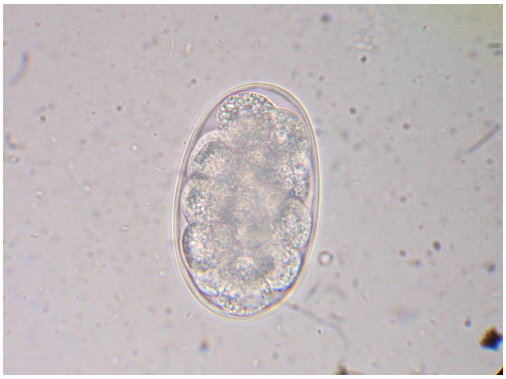


Figure 1: Oeufs d'Ankylostosme, par Joel Mills - CC BY-SA 3.0

Et par la suite, deviennent les des vers se propageant vers l'intestin:

Figure 2: Vers d'Ankylostosme, par CDC's Public Health Image Library

Voici le cycle parasitaire de l'ankylostome:

Nous nous intéressons donc aux facteurs favorisant l'infection des individus en cherchant les relations entre les variables présentes dans la table de données (âge, sexe, port de chaussures, zone géographique) et la présence ou non d'une infection. Dans le but d'identifier les variables les plus importantes et donc les facteurs influants, nous utiliserons principalement les modèles logistiques (simples et polytomiques ordonnées).

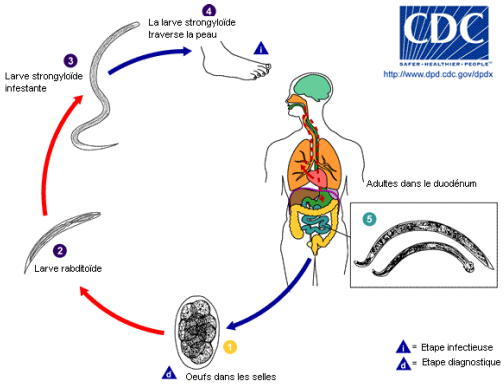


Figure 3: Cycle parasitaire de l'ankylostome, par CDC - Department of Parasitic Diseases - Domaine public

Quant aux données utilisées, il s'agit d'une enquête réalisée auprès de 637 individus avec diverses variables (nombre d'oeufs, intensité de la maladie, port de chaussures, ...) et nous allons créer une variable binaire *malade* afin de classer facilement les individus infectés.

Dans un premier temps, nous allons explorer les données (préparation, vérification des valeurs et quelques analyses descriptives) afin d'appréhender correctement nos données. Ensuite nous analyserons les relations entre les variables (test statistiques et visualisations graphiques). Dans un deuxième temps, nous mettrons en places des modèles de régression logistique pour comprendre l'influence des variables sur l'infection. Pour finir, nous sélectionnerons le meilleur modèle et évaluerons sa performance.

I. Lecture des données et vérification

A travers le tableau suivant, voici un récapitulatif de nos données:

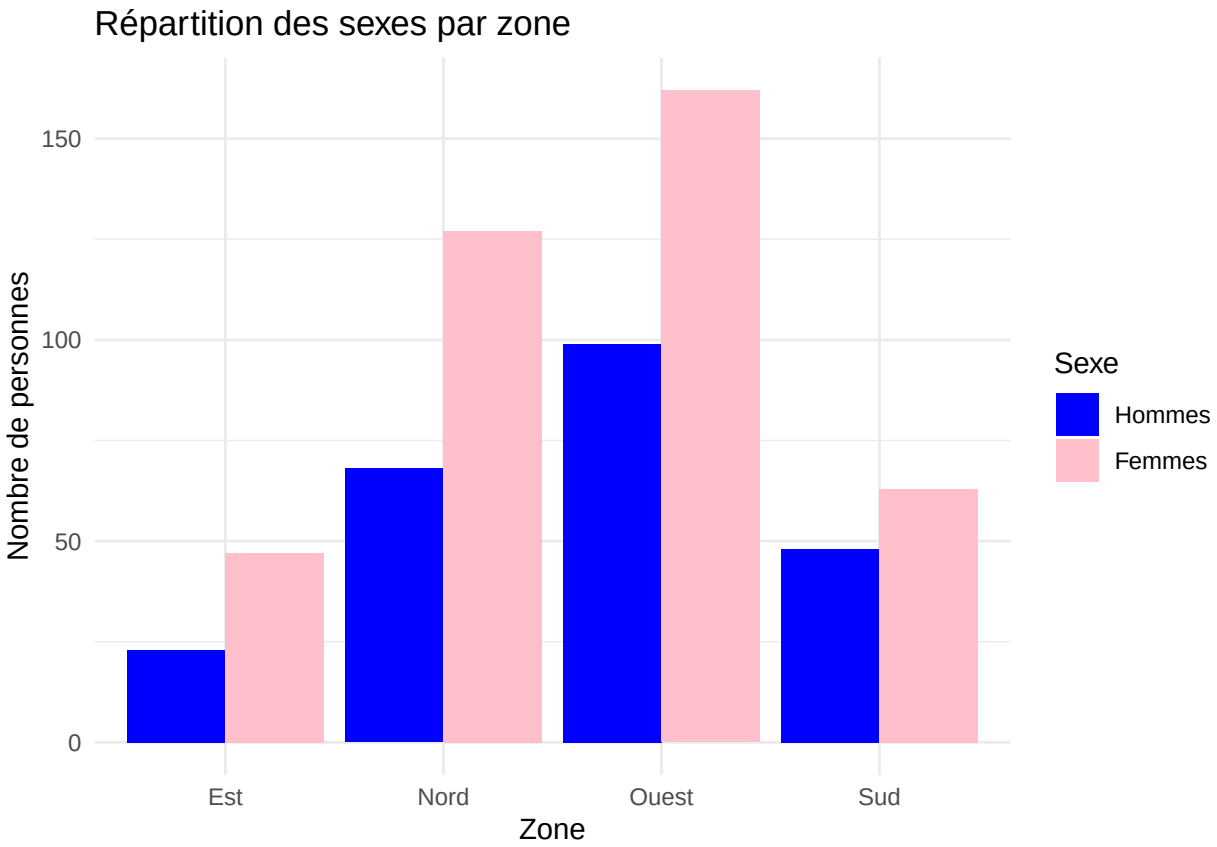
Nom de la variable	Type	Modalités ou exemples de modalités
id	int	440, 336, 60, ...
age	int	2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, ...
agegr	chr	<15 yrs, 15-59 yrs, 60+ yrs
zone	chr	Nord, Ouest, Sud, Est
sexe	int	0 (masculin ?), 1 (féminin ?)
chaussures	chr	"no", "yes"
nb.oeufs	int	0, 46, 184, 989, 1150, 690, ...
intensite	chr	0, "[1;1.999]", "[2;+]"
ageclasses	chr	<16 ans, 16-49, 49 et plus

Voici l'allure générale de nos données:

```
##    id age   agegr  zone sexe chaussures nb.oeufs   intensite ageclasses
## 1 440   2 <15 yrs  Nord    0         no        0           0    <16 ans
## 2 336   2 <15 yrs Ouest    1         no       46 "[1;1.999]"    <16 ans
## 3  60   2 <15 yrs  Nord    1         no      184 "[1;1.999]"    <16 ans
## 4 100   2 <15 yrs Ouest    1         no        0           0    <16 ans
## 5 281   2 <15 yrs   Sud    1         no        0           0    <16 ans
## 6  90   2 <15 yrs Ouest    0         no        0           0    <16 ans
```

I.1 Informations concernant les individus

La table de données contient 238 hommes et 399 femmes et ils sont répartis de la manière suivante selon la zone géographique:



Regardons les catégories d'âges. Trois variables sont à notre disposition: age, agegr et ageclasses.

Concernant la variable **age**:

Statistique	Valeur
Min.	2.00
1st Qu.	9.00
Median	23.00
Mean	25.94
3rd Qu.	40.00
Max.	78.00

Concernant la variable **agegr**:

Catégorie	Valeur
<15 yrs	259
15-59 yrs	331
60+ yrs	47

Concernant la variable **ageclasses**:

Catégorie	Valeur
<16 ans	259

Catégorie	Valeur
16-49 ans	331
49 et plus	47

I.2 Création de la variable “malade” et observations

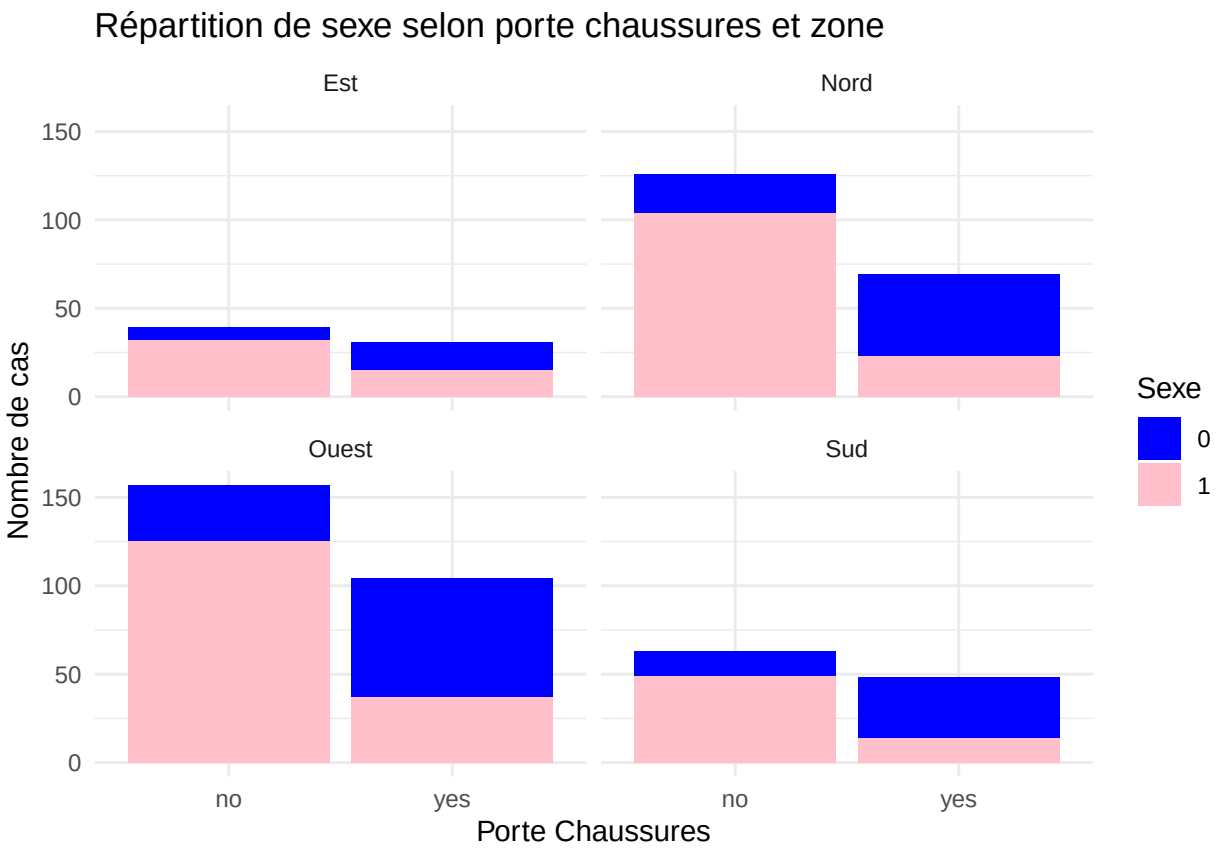
Afin de réaliser notre étude, nous créons la variable **malade**. Nous considérons qu'un individu est infecté si la variable **nb.oeufs** est supérieur à 0. La variable malade vaudra 0 si la variable **nb.oeufs** est égal à 0, sinon elle vaudra 1.

Suite à cette création nous constatons qu'il y a 197 personnes non malades (31% de l'échantillon) et 440 personnes malades (69% de l'échantillon), pour un total de 637 personnes.

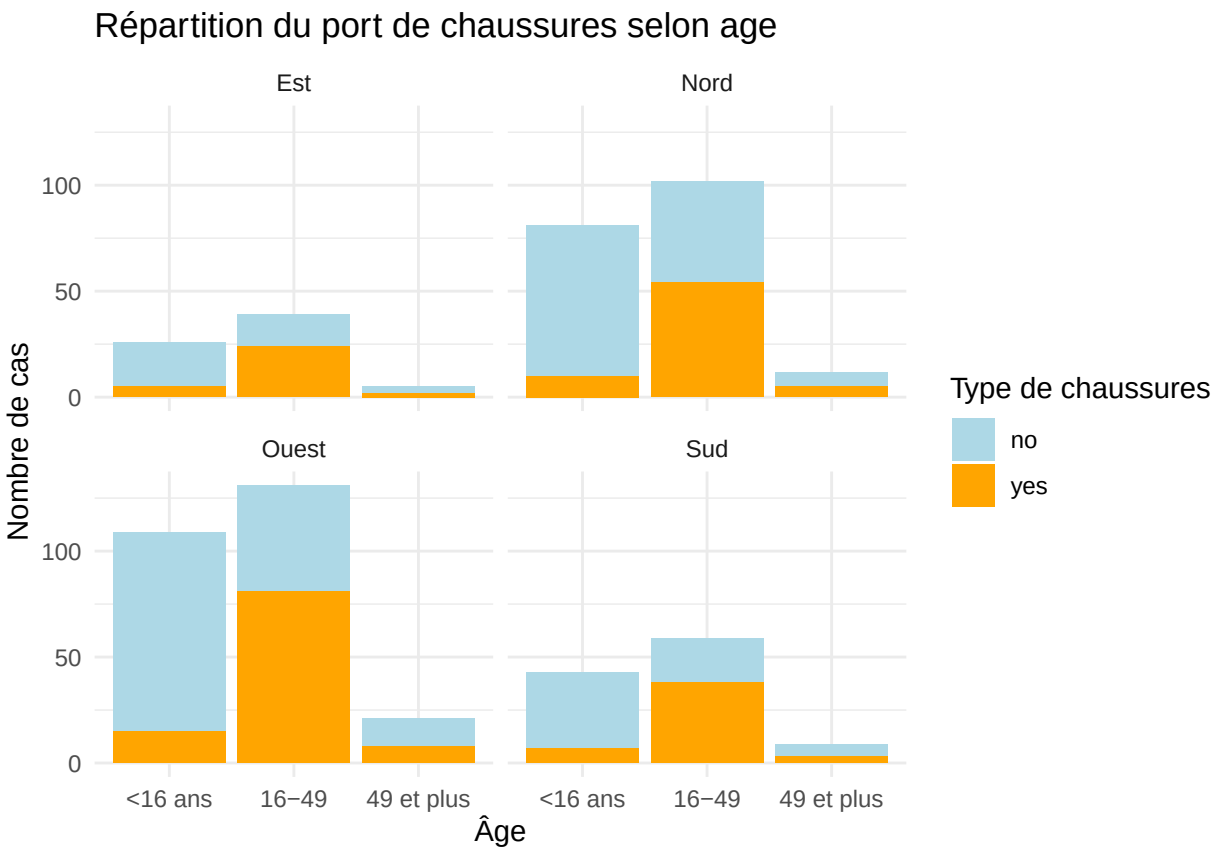
I.3 Quelques analyses préliminaires . . .

I.3.a . . . via des graphiques

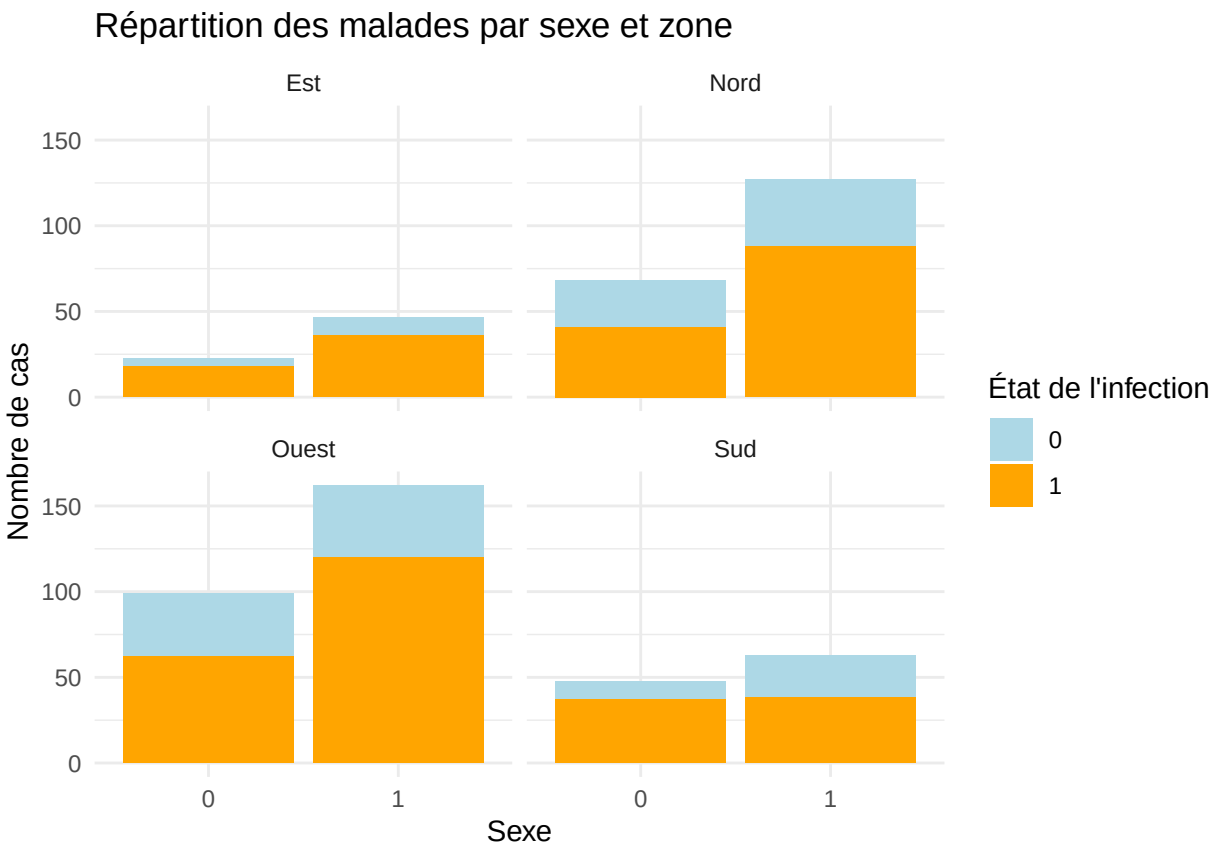
La répartition des sexes par zone géographique selon le porte de chaussures :



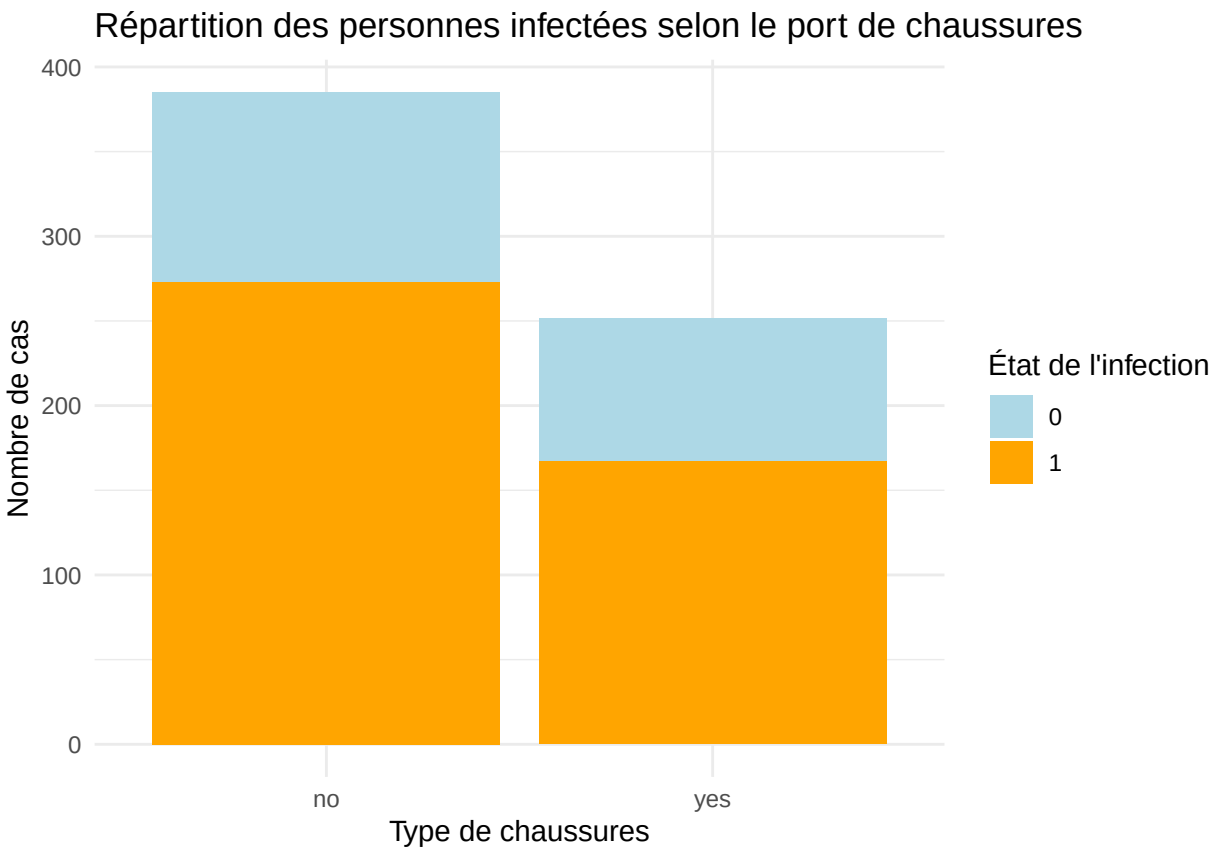
La répartition des personnes porte des chaussures selon l'âge par zone géographique :



La répartition des sexes par zone géographique selon l'infection:



La répartition des personnes infectées et le port de chaussures:

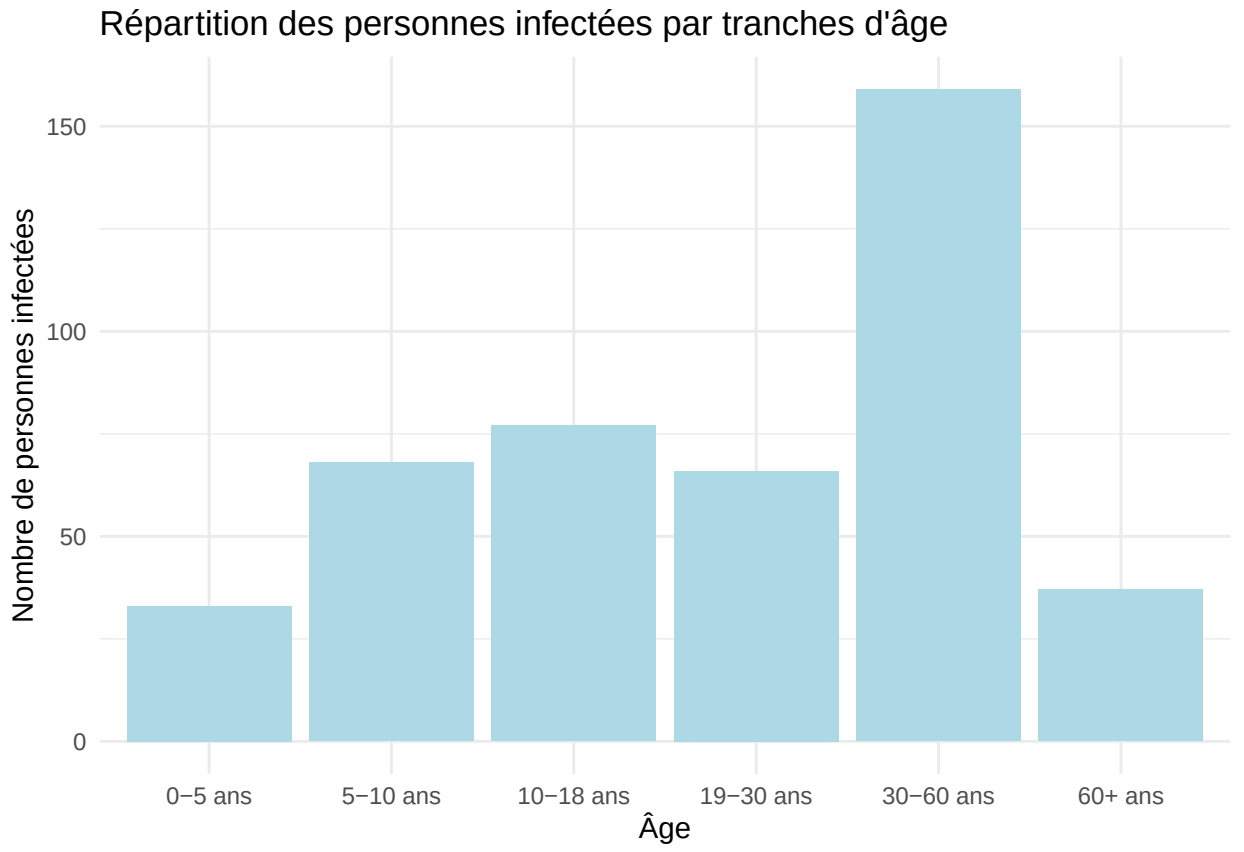


La répartition des personnes infectées selon leur âge. Pour cela nous allons créer une nouvelle variable age_categ car les deux autres variables qui classent les âges ne nous semblent pas représentatives. Voici les tranches choisies:
- 0 à 5 ans exclus,
- 5 ans à 10 ans exclus,
- 10 ans à 18 ans exclus,
- 19 ans à 30 ans exclus,
- 30 à 60 ans exlus,
- les plus de 60 ans

Test du khi-deux

Analyse des relations entre variables

variable_cible	variable_testee	p_valeur	interpretation
zone	chaussures	0.4378	Indépendance entre zone et chaussures
zone	ageclasses	0.9627	Indépendance entre zone et ageclasses
zone	sexe	0.4234	Indépendance entre zone et sexe
ageclasses	chaussures	0.0000	Dépendance entre ageclasses et chaussures
ageclasses	sexe	0.0000	Dépendance entre ageclasses et sexe
chaussures	sexe	0.0000	Dépendance entre chaussures et sexe



I.3.b . . . via le test du khi-deux

Le χ^2 permet de vérifier une relation entre deux variables qualitatives et de comparer des répartitions d'effectifs. Nous allons réaliser un text du khi-deux d'homogénéité et d'indépendance.

De plus, nous restons vigilants à la contrainte suivante : 80% des effectifs doivent être supérieurs à 5 individus.

Nous testons l'hypothèse nulle H0, les deux variables sont indépendantes contre H1, il existe une relation entre les deux variables testées.

Dans un premier temps, nous réaliserons ce test avec *zone* en tant que variable cible, avec *chaussures*, *age_categ* et *sexe*. Ensuite, nous réaliserons ce test entre *age_categ* et *chaussures*, *age_categ* et *sexe*, *chaussures* et *sexe*.

II. Modèles

II.1 Modèles logistiques avec une variable qualitative

La variable cible est *malade*, variable binaire où 1 représente une infection et 0 une absence d'infection.

Tout d'abord, nous allons regarder différents modèles logistiques selon seulement une variable qualitative: chaussures ; age ; sexe.

Nous choisirons dans chaque variable, la modalité de référence comme la modalité la plus réprésentée dans l'échantillon afin d'obtenir une diminution de la variance des estimateurs.

Variable chaussures

```
##
##  no yes
## 385 252
```

Les modalités possibles de la variable *chaussures* sont **no** ou **yes**. La modalité "no" représente 385 observations et "yes" 252 observations. Ainsi la modalité de référence déclarée sera "yes" et voici le modèle considéré :

$$malade = \beta_0 + \beta_{chaussuresno} * chaussuresno$$

```
### MODELE DE REGRESSION - VAR chaussures ###
res_chaussures <- glm(malade ~ chaussures, family="binomial", data=data)
res_chaussures
```

```
##
## Call:  glm(formula = malade ~ chaussures, family = "binomial", data = data)
##
## Coefficients:
##  (Intercept)  chaussuresno
##       0.6753        0.2156
##
## Degrees of Freedom: 636 Total (i.e. Null);  635 Residual
## Null Deviance:       788
## Residual Deviance: 786.5     AIC: 790.5
```

```
confint(res_chaussures, level=0.95)
```

```
## Waiting for profiling to be done...
```

```
##                    2.5 %    97.5 %
## (Intercept)   0.4175228 0.9405830
## chaussuresno -0.1268667 0.5565847
```

Nous obtenons $\beta_{chaussuresno} = 0.2156$ et $IC_{95\%} = [-0.1268667, 0.5565847]$. Le groupe de référence est représenté par la modalité *yes*. Ainsi la valeur obtenue pour $\beta_{chaussuresno} - \beta_{chaussuresyes} = ln(\frac{odds(malade=1|chaussures="no")}{odds(malade=1|chaussures="yes")})$ est 0.2156, soit :

$$\frac{odds(malade = 1|chaussures = "no")}{odds(malade = 1|chaussures = "yes")} = \exp(0.2156) = 1.24$$

Nous pouvons dire que la sous-population des personnes ne portant pas de chaussures a 1.24 fois plus de chance d'infection que la sous-population portant des chaussures. L'intervalle de confiance obtenue contient 0 et est large, il est difficile de conclure en raison de son imprécision.

Variable sexe

```
##
##   0   1
## 238 399
```

Les modalités possibles de la variable *sexe* sont **O** ou **1**. La modalité "0" (homme) représente 238 observations et "1"(femme) 399 observations. Ainsi la modalité de référence déclarée sera "1" et voici le modèle considéré :

$$malade = \beta_0 + \beta_{sexe0} * sexe0$$

```
### MODELE DE REGRESSION - VAR chaussures ###
res_sexe <- glm(malade ~ sexe, family="binomial", data=data)
res_sexe
```

```
##
## Call:  glm(formula = malade ~ sexe, family = "binomial", data = data)
##
## Coefficients:
## (Intercept)        sexe0
##      0.8797      -0.1992
##
## Degrees of Freedom: 636 Total (i.e. Null);  635 Residual
## Null Deviance:       788
## Residual Deviance: 786.7     AIC: 790.7
```

```
confint(res_sexe, level=0.95)
```

```
## Waiting for profiling to be done...
```

```
##                  2.5 %    97.5 %
## (Intercept)  0.6671876 1.0986994
## sexe0       -0.5429823 0.1469779
```

Nous obtenons $\beta_{sexe0} = 0.1992$ et $IC_{95\%} = [-0.5429823, 0.1469779]$. Le groupe de référence est représenté par la modalité "1"(femme). Ainsi la valeur obtenue pour $\beta_{sexe0} - \beta_{sexe1} = ln(\frac{odds(malade=1|sexe="0")}{odds(malade=1|sexe="1")})$ est 0.2156, soit :

$$\frac{odds(malade = 1|sexe = "0")}{odds(malade = 1|sexe = "1")} = \exp(-0.1992) = 0.82$$

Nous pouvons dire que la sous-population des hommes a 0.82 fois plus de chance d'infection que la sous-population représentant les femmes, soit que les hommes ont moins de chance d'être infectés. L'intervalle de confiance obtenue contient 0 et est large, il est difficile de conclure en raison de son imprécision.

Variable age

```
## Modèle GLM avec la variable age
res_age <- glm(malade ~ age_categ, family="binomial", data=data)
res_age
```

```
##
## Call:  glm(formula = malade ~ age_categ, family = "binomial", data = data)
##
## Coefficients:
##        (Intercept)   age_categ5-10 ans  age_categ10-18 ans  age_categ19-30 ans
##            -0.3102              1.0033              1.3218              1.3218
## age_categ30-60 ans    age_categ60+ ans
##             1.3537              1.6185
##
## Degrees of Freedom: 636 Total (i.e. Null);  631 Residual
## Null Deviance:       788
## Residual Deviance: 757.6     AIC: 769.6
```

```
confint(res_age, level=0.95)
```

```
## Waiting for profiling to be done...
```

```
##                          2.5 %    97.5 %
## (Intercept)        -0.7664682 0.1357438
## age_categ5-10 ans   0.3998837 1.6209862
## age_categ10-18 ans  0.7064985 1.9564520
## age_categ19-30 ans  0.6834113 1.9828250
## age_categ30-60 ans  0.8156896 1.9035118
## age_categ60+ ans    0.8172410 2.4900726
```

On voit que toutes les covariables ne contiennent pas 0 dans leur intervalle de confiance.

$$\frac{odds(malade = 1|age\_categ = 5 - 10\ ans)}{odds(malade = 1|age\_categ = 0 - 4\ ans)} \geqslant exp(0.40) \approx 1.49$$

Les individus de 5-10 ans ont plus de 1.49 fois plus de chance d'être malade que ceux de 0-4 ans

$$\frac{odds(malade = 1|age\_categ = 10 - 18\ ans)}{odds(malade = 1|age\_categ = 0 - 4\ ans)} \geqslant exp(0.7) \approx 1.82$$

Les individus de 10-18 ans ont plus de 1.82 fois plus de chance d'être malade que ceux de 0-4 ans

$$\frac{odds(malade = 1|age\_categ = 19 - 30\ ans)}{odds(malade = 1|age\_categ = 0 - 4\ ans)} \geqslant exp(0.68) \approx 1.97$$

Les individus de 19-30 ans ont plus de 1.97 fois plus de chance d'être malade que ceux de 0-4 ans

$$\frac{odds(malade = 1|age\_categ = 31 - 60\ ans)}{odds(malade = 1|age\_categ = 0 - 4\ ans)} \geqslant exp(0.82) \approx 2.27$$

Les individus de 31-60 ans ont plus de 2.27 fois plus de chance d'être malade que ceux de 0-4 ans

$$\frac{odds(malade = 1|age\_categ = 60 + \ ans)}{odds(malade = 1|age\_categ = 0 - 4 \ ans)} \geqslant exp(0.82) \approx 2.27$$

Les individus de 60 ans et plus ont plus de 2.27 fois plus de chance d'être malade que ceux de 0-4 ans

II.2 Modèles logistiques avec plus d'une variable qualitative

Relations entre les variables :

- chaussures et ageclasses : p = 0.0000

Il y a une relation significative entre le port de chaussures et les classes d'âge. Cela suggère que l'effet du port de chaussures sur la probabilité d'être malade pourrait dépendre de l'âge.

- ageclasses et sexe : p = 0.0000

Il y a une relation significative entre les classes d'âge et le sexe. Cela signifie que la distribution des classes d'âge varie selon le sexe.

- chaussures et sexe : p = 0.0000

Il y a une relation significative entre le port de chaussures et le sexe. Cela suggère que le port de chaussures pourrait varier selon le sexe.

=> Ces relations significatives indiquent que :

Les variables chaussures, ageclasses et sexe ne sont pas indépendantes.

Il est important de considérer les interactions entre ces variables dans le modèle GLM pour capturer leurs effets combinés.

Option 1 : Modèle avec interactions entre port de chaussures et l'âge :

malade $= \beta_0 + \beta_{chaussuresno}\text{chaussuresno} + \beta_{16-49}\text{ageclasses}_{16-49} + \beta_{49+}\text{ageclasses}_{49+} + \beta_1(\text{chaussuresno} \times \text{ageclasses}_{16-49}) + \beta_2($

```
res_cha.age <- glm(malade ~ chaussures*ageclasses, family="binomial", data=data)
res_cha.age
```

```
##
## Call:  glm(formula = malade ~ chaussures * ageclasses, family = "binomial",
##     data = data)
##
## Coefficients:
##                         (Intercept)                         chaussuresno
##                             0.61310                             -0.19260
##                    ageclasses16-49                ageclasses49 et plus
##                             0.07244                              0.08004
##      chaussuresno:ageclasses16-49  chaussuresno:ageclasses49 et plus
##                             1.24753                              1.33204
##
## Degrees of Freedom: 636 Total (i.e. Null);  631 Residual
## Null Deviance:       788
## Residual Deviance: 756.5     AIC: 768.5
```

```r
confint(res_cha.age, level=0.95)
```

```
## Waiting for profiling to be done...
```

```
##                                       2.5 %    97.5 %
## (Intercept)                      -0.04562779 1.3175075
## chaussuresno                     -0.94419508 0.5200309
## ageclasses16-49                  -0.68845103 0.7962925
## ageclasses49 et plus             -1.09433528 1.3154381
## chaussuresno:ageclasses16-49      0.34583228 2.1867747
## chaussuresno:ageclasses49 et plus -0.26044977 3.0114739
```

Coefficients du modèle Les coefficients indiquent l'effet de chaque variable sur la probabilité d'être malade (en log-odds). Une valeur positive augmente la probabilité, tandis qu'une valeur négative la diminue.

- **(Intercept) :** 0.61310 C'est la valeur de référence (log-odds) lorsque toutes les variables sont à leur niveau de référence (chaussures = yes, ageclasses = moins de 16 ans).
- **chaussuresno :** -0.19260 Le fait de ne pas porter de chaussures (chaussures = no) diminue légèrement les log-odds d'être malade par rapport au port de chaussures, mais cet effet n'est pas significatif (l'intervalle de confiance à 95 % inclut 0).
- **ageclasses16-49 :** 0.07244 Les personnes âgées de 16 à 49 ans ont des log-odds légèrement plus élevés d'être malades que celles de moins de 16 ans, mais cet effet n'est pas significatif.
- **ageclasses49 et plus :** 0.08004 Les personnes de 49 ans et plus ont des log-odds légèrement plus élevés d'être malades que celles de moins de 16 ans, mais cet effet n'est pas significatif.
- **chaussuresno:ageclasses16-49 :** 1.24753 Il y a une interaction positive et significative entre le fait de ne pas porter de chaussures et la classe d'âge 16-49 ans. Cela signifie que, pour cette tranche d'âge, ne pas porter de chaussures augmente significativement les log-odds d'être malade.
- **chaussuresno:ageclasses49 et plus :** 1.33204

Il y a également une interaction positive entre le fait de ne pas porter de chaussures et la classe d'âge 49 ans et plus., mais cet effet n'est pas significatif.

Conclusion du modèle L'interaction entre chaussuresno et ageclasses16-49 est significative, ce qui suggère que l'effet du port de chaussures sur la probabilité d'être malade dépend de l'âge pour cette tranche d'âge.

Pour ageclasses49 et plus, l'interaction n'est pas significative, mais la valeur estimée est élevée (1.33204) avec un intervalle de confiance large. Cela pourrait indiquer un manque de puissance statistique due à un échantillon insuffisant dans cette tranche d'âge.

Un modèle simplifié pourrait inclure uniquement l'interaction significative (chaussuresno:ageclasses16-49)

Option 2 : Modèle avec interactions entre port de chaussures et le sexe :

$$\text{malade} = \beta_0 + \beta_{\text{chaussuresno}} \cdot \text{chaussuresno} + \beta_{\text{sexe0}} \cdot \text{Sexe}_0 + \beta_{\text{interaction}} \cdot (\text{chaussuresno} \times \text{Sexe}_0)$$

```r
res_cha.sex <- glm(malade ~ chaussures*sexe, family="binomial", data=data)
res_cha.sex
```

```
##
## Call:  glm(formula = malade ~ chaussures * sexe, family = "binomial",
##     data = data)
##
## Coefficients:
##        (Intercept)        chaussuresno               sexe0  chaussuresno:sexe0
##            0.83130             0.06252            -0.23778             0.22321
##
## Degrees of Freedom: 636 Total (i.e. Null);  633 Residual
## Null Deviance:       788
## Residual Deviance: 785.7     AIC: 793.7
```

```
confint(res_cha.sex, level=0.95)
```

```
## Waiting for profiling to be done...
```

```
##                          2.5 %    97.5 %
## (Intercept)          0.3908125 1.2987157
## chaussuresno        -0.4627775 0.5688407
## sexe0               -0.8007575 0.3105208
## chaussuresno:sexe0 -0.5517709 1.0184625
```

Coefficients du modèle

- **(Intercept) :** 0.83130 C'est la valeur de référence (log-odds) lorsque toutes les variables sont à leur niveau de référence(chaussures = yes,sexe = 1 les hommes).)
- **chaussuresno :** 0.06252 Le fait de ne pas porter de chaussures (chaussures = no) augmente légèrement les log-odds d'être malade par rapport au port de chaussures, mais cet effet n'est pas significatif (l'intervalle de confiance à 95 % inclut 0).
- **sexe0 :** -0.23778

Le sexe de référence sexe = 0 (les femmes) a des log-odds légèrement plus faibles d'être malade que le sexe de référence sexe = 1, mais cet effet n'est pas significatif.

- **chaussuresno:sexe0 :** 0.22321

Il y a une interaction positive entre le fait de ne pas porter de chaussures et le sexe sexe = 0. Cela signifie que, pour le sexe sexe = 0, ne pas porter de chaussures augmente légèrement les log-odds d'être malade. Cependant, cet effet n'est pas significatif.

Conclusion du modèle le modèle n'identifie aucun effet significatif des variables explicatives (chaussures, sexe et leur interaction) sur la variable cible (malade)

Option 3 : Modèle avec interactions entre port de chaussures et sexe avec l'effet additif d'âge :

malade = $\beta_0 + \beta_{\text{chaussuresno}} \cdot \text{chaussuresno} + \beta_{\text{sexe0}} \cdot \text{Sexe}_0 + \beta_{\text{ageclasses16-49}} \cdot \text{ageclasses16-49} + \beta_{\text{ageclasses49 et plus}} \cdot \text{ageclasses49 et plus}$

```r
res_cha.age_sex <- glm(malade ~ chaussures*ageclasses + sexe, family="binomial", data=data)
res_cha.age_sex
```

```
##
## Call:  glm(formula = malade ~ chaussures * ageclasses + sexe, family = "binomial",
##     data = data)
##
## Coefficients:
##                          (Intercept)                         chaussuresno
##                              0.72396                           -0.27201
##                    ageclasses16-49                ageclasses49 et plus
##                              0.08015                            0.05876
##                                sexe0       chaussuresno:ageclasses16-49
##                             -0.17655                            1.25254
## chaussuresno:ageclasses49 et plus
##                              1.34759
##
## Degrees of Freedom: 636 Total (i.e. Null);  630 Residual
## Null Deviance:       788
## Residual Deviance: 755.7     AIC: 769.7
```

```r
confint(res_cha.age_sex, level=0.95)
```

```
## Waiting for profiling to be done...
```

```
##                                       2.5 %    97.5 %
## (Intercept)                        0.01824004 1.4726839
## chaussuresno                      -1.04575335 0.4634776
## ageclasses16-49                   -0.68143990 0.8049178
## ageclasses49 et plus              -1.11782714 1.2958217
## sexe0                             -0.57614423 0.2236899
## chaussuresno:ageclasses16-49       0.35012245 2.1925361
## chaussuresno:ageclasses49 et plus -0.24618336 3.0283592
```

Coefficients du modèle

- **ageclasses16-49 :** 0.91803

Les personnes âgées de 16 à 49 ans ont des log-odds significativement plus élevés d'être malades que celles de moins de 16 ans (la classe de référence).

- **ageclasses49 et plus :** 1.02286

Les personnes de 49 ans et plus ont des log-odds significativement plus élevés d'être malades que celles de moins de 16 ans (la classe de référence).

- Les autre coefficient sont non significatifs.

Modèle	AIC
cha * sexe + age	775.6
cha * age + sexe	769.7
sexe * age + chau	776.9
cha * age	768.5
cha * sexe	793.7

Modèle polytomique ordonné

```
echelle_maladie <- function(x){
  if (x == 0){
    "pas malade"
  }else if (x>0 & x < 2000){
    "peu malade"
  }else{
    "très malade"
  }
}
data$malade_echelle <- sapply(data$nb.oeufs,FUN = echelle_maladie)
```

```
library(MASS)
```

```
##
## Attachement du package : 'MASS'

## L'objet suivant est masqué depuis 'package:dplyr':
##
##     select
```

```
data$malade_echelle <- as.factor(data$malade_echelle)
res_polyt <- polr(malade_echelle~sexe ,data = data)
levels(data$malade_echelle)
```

```
## [1] "pas malade"  "peu malade"  "très malade"
```

```
res_polyt
```

```
## Call:
## polr(formula = malade_echelle ~ sexe, data = data)
##
## Coefficients:
##      sexe0
## -0.2915572
##
## Intercepts:
##  pas malade|peu malade peu malade|très malade
##             -0.9182877              1.6878877
##
## Residual Deviance: 1233.117
## AIC: 1239.117
```

Comparaison des modèles

Modèls	AIC
cha* sexe + age	775.6
cha* age + sexe	769.7
sexe*age +chau	776.9
cha* age	768.5
cha* sexe	793.7

Selon le critère AIC, le meilleur modèle pour prédire la maladie est celui incluant l'interaction entre chaussures et âge (cha*age), car il présente la plus faible valeur d'AIC (768.5).

III. Prédictions

```
### PREDICTIONS ###
# vect_estimations <- round(res$fitted.values)
#
# #Effectif
# tab=table(data$malade, vect_estimations)
# tab
#
# #Proportion de personnes pour laquelle la prédiction a été mauvaise: 197 (1 + 196)
# #1. => 31%
# (tab[1,2] + tab[2,1])
# (tab[1,2] + tab[2,1])/sum(tab)
#
# #2. Proportion de personnes infectées pour laquelle la prédiction était non infecté: 31% (faux positi
# tab[1,2]/sum(tab[,2])
#
# #3. Proportion de personnes non infectées pour laquelle la prédiction était infectées => 50% (faux né
# tab[2,1]/sum(tab[,1])
```

Conclusion

A MODIFIER

Il est important de noter que pour prévenir la population de ce type d'infection, il vaut mieux éviter de marcher pieds nus, d'utiliser des eaux usées et de bien utiliser des dispositifs de toilettes, d'hygiène pour éviter la présence de selles au sol. Le diagnostic de l'infection peut-être réalisé via un examen d'un échantillon de selles ou d'analyse de sang.

Annexe Code R

```r
knitr::opts_chunk$set(echo = TRUE)
### LIBRAIRIES UTILISEES ###
library(dplyr)
library(ggplot2)
library(gt)
### LECTURE DES DONNEES ET MODALITES ###
data <- read.csv("Ankylostome.csv")
data <- data %>% select(-c(...1, X))

modalites_uniques <- lapply(data, function(colonne) {
  unique_values <- unique(colonne)
  count_values <- length(unique_values)
  list(Modalites = unique_values, Nombre = count_values)
})

### ALLURE GENERALE DES DONNÉES ###
head(data)

### SEXE DES INDIVIDUS ###
table_sexe <- table(data$sexe)

### REPARTITION SELON LES ZONES ###
ggplot(data, aes(x = zone, fill = as.factor(sexe))) +
  geom_bar(position = "dodge") +
  labs(
    title = "Répartition des sexes par zone",
    x = "Zone",
    y = "Nombre de personnes",
    fill = "Sexe"
  ) +
  scale_fill_manual(
    values = c("0" = "blue", "1" = "pink"),
    labels = c("0" = "Hommes", "1" = "Femmes")
  ) +
  theme_minimal()

### AGE DES INDIVIDUS ###
age <- summary(data["age"])
agegr <- table(data$agegr)
ageclasses <- table(data$ageclasses)

### CREATION DE LA VARIABLE MALADE ET OBSERVATIONS ###
data <- data %>% mutate(malade = ifelse(nb.oeufs == 0, 0, 1))
malades <- table(data$malade)
pourcentages_malades <- prop.table(malades) * 100

### répartition d'age par zone géographique selon la variable Sexe ###
ggplot(data, aes(x = as.factor(chaussures), fill = as.factor(sexe))) +
  geom_bar(position = "stack") +
  facet_wrap(~ zone) +
  labs(title = "Répartition de sexe selon porte chaussures et zone",
```

```
       x = "Porte Chaussures",
       y = "Nombre de cas",
       fill = "Sexe") +
  scale_fill_manual(values = c("blue", "pink")) +
  theme_minimal()

### répartition du port de chaussures par rapport au classe d'age ###
ggplot(data, aes(x = ageclasses, fill = chaussures)) +
  geom_bar(position = "stack") +
  facet_wrap(~ zone) +
  labs(title = "Répartition du port de chaussures selon age",
       x = "Âge",
       y = "Nombre de cas",
       fill = "Type de chaussures") +
  scale_fill_manual(values = c("lightblue", "orange")) +
  theme_minimal()

### répartition des sexes par zone géographique selon l'infection ###
ggplot(data, aes(x = as.factor(sexe), fill = as.factor(malade))) +
  geom_bar(position = "stack") +
  facet_wrap(~ zone) +
  labs(title = "Répartition des malades par sexe et zone",
       x = "Sexe",
       y = "Nombre de cas",
       fill = "État de l'infection") +
  scale_fill_manual(values = c("lightblue", "orange")) +
  theme_minimal()

### répartition des personnes infectées et le port de chaussures ###
ggplot(data, aes(x = chaussures, fill = as.factor(malade))) +
  geom_bar(position = "stack") +
  labs(title = "Répartition des personnes infectées selon le port de chaussures",
       x = "Type de chaussures",
       y = "Nombre de cas",
       fill = "État de l'infection") +
  scale_fill_manual(values = c("lightblue", "orange")) +
  theme_minimal()

### CREATION DES CLASSES D'AGE ET GRAPHIQUE ###
data$age_categ <- cut(data$age,
                      breaks = c(0, 5, 10, 18, 30, 60, Inf),
                      labels = c("0-5 ans", "5-10 ans", "10-18 ans", "19-30 ans", "30-60 ans", "60+ ans
                      right = FALSE)

data_malade <- subset(data, malade == 1)

ggplot(data_malade, aes(x = age_categ)) +
  geom_bar(fill = "lightblue") +
  labs(title = "Répartition des personnes infectées par tranches d'âge",
       x = "Âge",
       y = "Nombre de personnes infectées") +
  theme_minimal()
```

```r
resultat <- data.frame(
  variable_cible = character(),
  variable_testee = character(),
  p_valeur = numeric(),
  interpretation = character())

tests <- list(
  list(cible = "zone", testee = "chaussures"),
  list(cible = "zone", testee = "ageclasses"),
  list(cible = "zone", testee = "sexe"),
  list(cible = "ageclasses", testee = "chaussures"),
  list(cible = "ageclasses", testee = "sexe"),
  list(cible = "chaussures", testee = "sexe")
)

for (test in tests) {
  cible <- test$cible
  testee <- test$testee
    table_contingence <- table(data[[cible]], data[[testee]])

  if (any(chisq.test(table_contingence)$expected < 5)) {
    p_valeur <- NA
    interpretation <- "Test invalide (effectifs attendus < 5)"
  } else {
    test_resultat <- chisq.test(table_contingence)
    p_valeur <- test_resultat$p.value
    interpretation <- ifelse(
      p_valeur < 0.05,
      paste("Dépendance entre", cible, "et", testee),
      paste("Indépendance entre", cible, "et", testee)
    )
  }

  resultat <- rbind(
    resultat,
    data.frame(
      variable_cible = cible,
      variable_testee = testee,
      p_valeur = round(p_valeur, 4),
      interpretation = interpretation
    )
  )
}

resultat %>%
  gt() %>%
  tab_header(
    title = "Test du khi-deux",
    subtitle = "Analyse des relations entre variables"
  )

table(data$chaussures)
```

```r
data$chaussures <- relevel(factor(data$chaussures), ref = "yes")
### MODELE DE REGRESSION - VAR chaussures ###
res_chaussures <- glm(malade ~ chaussures, family="binomial", data=data)
res_chaussures
confint(res_chaussures, level=0.95)
table(data$sexe)

data$sexe <- relevel(factor(data$sexe), ref = "1")
### MODELE DE REGRESSION - VAR chaussures ###
res_sexe <- glm(malade ~ sexe, family="binomial", data=data)
res_sexe
confint(res_sexe, level=0.95)

## Modèle GLM avec la variable age
res_age <- glm(malade ~ age_categ, family="binomial", data=data)
res_age
confint(res_age, level=0.95)
res_cha.age <- glm(malade ~ chaussures*ageclasses, family="binomial", data=data)
res_cha.age
confint(res_cha.age, level=0.95)
res_cha.sex <- glm(malade ~ chaussures*sexe, family="binomial", data=data)
res_cha.sex
confint(res_cha.sex, level=0.95)
res_cha.age_sex <- glm(malade ~ chaussures*ageclasses + sexe, family="binomial", data=data)
res_cha.age_sex
confint(res_cha.age_sex, level=0.95)
echelle_maladie <- function(x){
  if (x == 0){
    "pas malade"
  }else if (x>0 & x < 2000){
    "peu malade"
  }else{
    "très malade"
  }
}
data$malade_echelle <- sapply(data$nb.oeufs,FUN = echelle_maladie)
library(MASS)
data$malade_echelle <- as.factor(data$malade_echelle)
res_polyt <- polr(malade_echelle~sexe ,data = data)
levels(data$malade_echelle)
res_polyt
### PREDICTIONS ###
# vect_estimations <- round(res$fitted.values)
#
# #Effectif
# tab=table(data$malade, vect_estimations)
# tab
#
# #Proportion de personnes pour laquelle la prédiction a été mauvaise: 197 (1 + 196)
# #1. => 31%
# (tab[1,2] + tab[2,1])
# (tab[1,2] + tab[2,1])/sum(tab)
#
```

```
# #2. Proportion de personnes infectées pour laquelle la prédiction était non infecté: 31% (faux positi
# tab[1,2]/sum(tab[,2])
#
# #3. Proportion de personnes non infectées pour laquelle la prédiction était infectées => 50% (faux né
# tab[2,1]/sum(tab[,1])
```