Données de comptage avec excès de zéro

Modèles Zero-inflated

Séminaire de biostatistiques

Camille ALLEYRAT

09/01/2024

CHU Poitiers

CENTRE HOSPITALIER UNIVERSITAIRE

Plan

1) Données de comptage et excès de zéro
2) Zero-inflated Poisson
3) Zero-inflated Negative Binomial
4) Lequel choisir?

Contexte - donnée de comptage

Donnée de comptage = variable résultant d'un processus de comptage

- Nombre d'œufs pondus par semaine
- Nombre de buts marqués pendant un match

→ Distribuées selon une loi de Poisson

un paramètre : λ = moyenne = variance

Contexte - modélisation

Modèle linéaire pas adéquat car homoscédasticité non respectée,

variance proportionnelle à la moyenne

$$\mathbb{E}(X) = Var(X) = \lambda$$

→ Régression de Poisson (modèle linéaire généralisé, GLM)

Lien log entre le prédicteur linéaire η et la réponse moyenne μ

$$\log(\mu) = \eta = \sum_{j=1}^{J} \beta_j X_{ij}$$

Contexte - excès de zéro (1)

Mais certaines données de comptage ne sont pas bien représentées par une distribution de Poisson

- Nombre de jours sans ventilation (*ventilator-free days*, VFD)
  - VFD = 0 si patient DCD avant J28
  - VFD = 28 – x sinon

    avec x = nombre de jours où le patient n'est plus sous ventilation entre J1 et J28

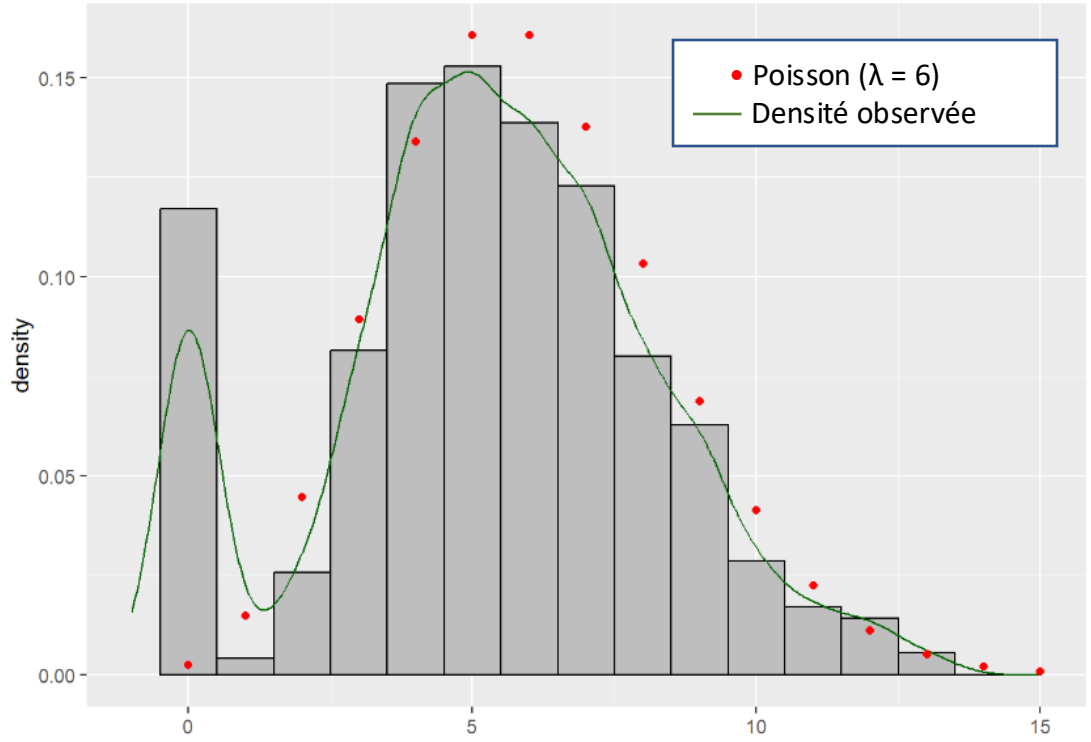
Contexte - excès de zéro (2)

- Nombre de jours sans ventilation (ventilator-free days, VFD)

  - VFD = 0 si patient DCD avant J28
  - VFD = 28 – x sinon

    avec x = nb jours où le patient n'est plus sous ventilation entre J1 et J28

→ Par définition, excès de zéro



→ Régression de Poisson n'est plus adaptée

→ **Modèles inflatés en zéro**

Modèles inflatés en zéro

1er exemple : Zero-inflated Poisson

Zero-inflated Poisson (ZIP)

Mélange de deux processus :

- Processus de comptage et processus de Bernoulli

$$P \sim \mathcal{P}(\lambda) \qquad Z \sim \mathcal{B}(\pi)$$

- Variable composite Y = (1-Z) + Z*P

- Processus de Bernoulli tient compte des **zéro surajoutés** par rapport à la distribution de comptage initiale

ZIP model – mise en œuvre (1)

Variable composite Y = (1-Z) + Z*P

Analysée avec un mélange de deux modèles :

- Régression de Poisson et régression logistique

$$log(\lambda) = \beta_0 + \sum_{j=1}^{J} \beta_j X_{ij} \qquad \text{(ZIP1)}$$

$$ln(\frac{\pi}{1-\pi}) = \gamma_0 + \sum_{j=1}^{J} \gamma_j X_{ij} \qquad \text{(ZIP2)}$$

- Avec R : fonction **`zeroinfl`** du package **`pscl`**

ZIP model – mise en œuvre (2)

Zero-inflated Poisson

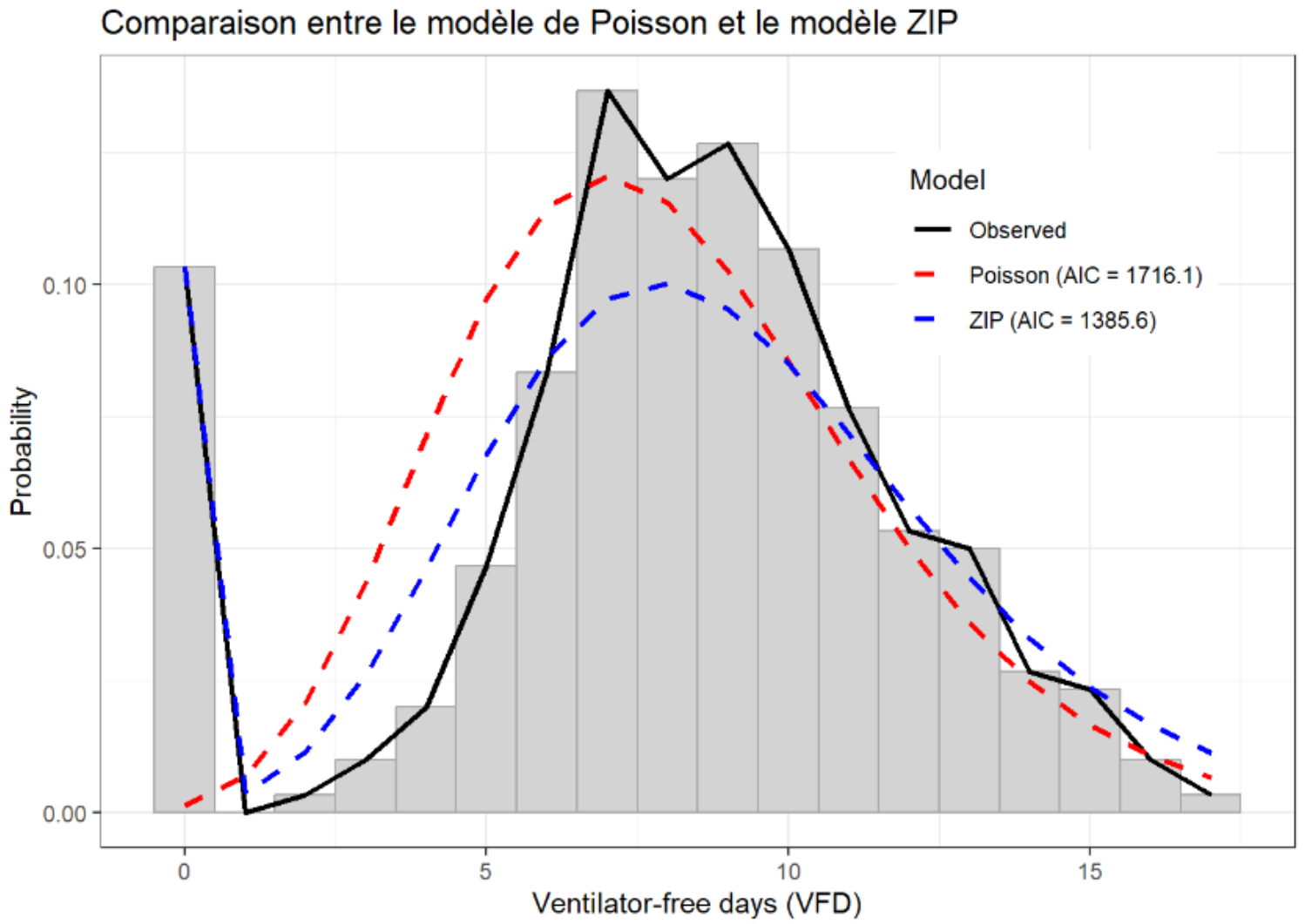
```
##
## Call:
## zeroinfl(formula = VFD ~ AGE + SEXE + RND | AGE + SEXE + RND, data = simu,
##     dist = "poisson")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -2.7147 -0.2385  0.1346  0.5280  1.4431
##
## Count model coefficients (poisson with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.010391   0.108157   9.342   <2e-16 ***
## AGE          0.030928   0.002649  11.677   <2e-16 ***
## SEXEH       -0.002836   0.040871  -0.069    0.945
## RNDB        -0.016045   0.042130  -0.381    0.703
##
## Zero-inflation model coefficients (binomial with logit link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.945634   1.000678  -2.944  0.00324 **
## AGE          0.009264   0.024912   0.372  0.70999
## SEXEH       -0.318913   0.397055  -0.803  0.42186
## RNDB         1.102250   0.398914   2.763  0.00573 **
##
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 11
## Log-likelihood: -684.8 on 8 Df
```

Poisson

```
##
## Call:
## glm(formula = VFD ~ RND + AGE + SEXE, family = poisson, data = simu)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.2032  -0.3218   0.2143   0.7308   2.1230
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.998302   0.106696   9.356  < 2e-16 ***
## RNDB        -0.126669   0.042049  -3.012  0.00259 **
## AGE          0.029229   0.002615  11.178  < 2e-16 ***
## SEXEH        0.021867   0.040842   0.535  0.59237
##
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 765.79  on 299  degrees of freedom
## Residual deviance: 632.05  on 296  degrees of freedom
## AIC: 1716.1
##
## Number of Fisher Scoring iterations: 5
```

ZIP model – comparaison avec Poisson

Le modèle ZIP permet de bien rendre compte de l'excès de zéro dans la distribution observée.



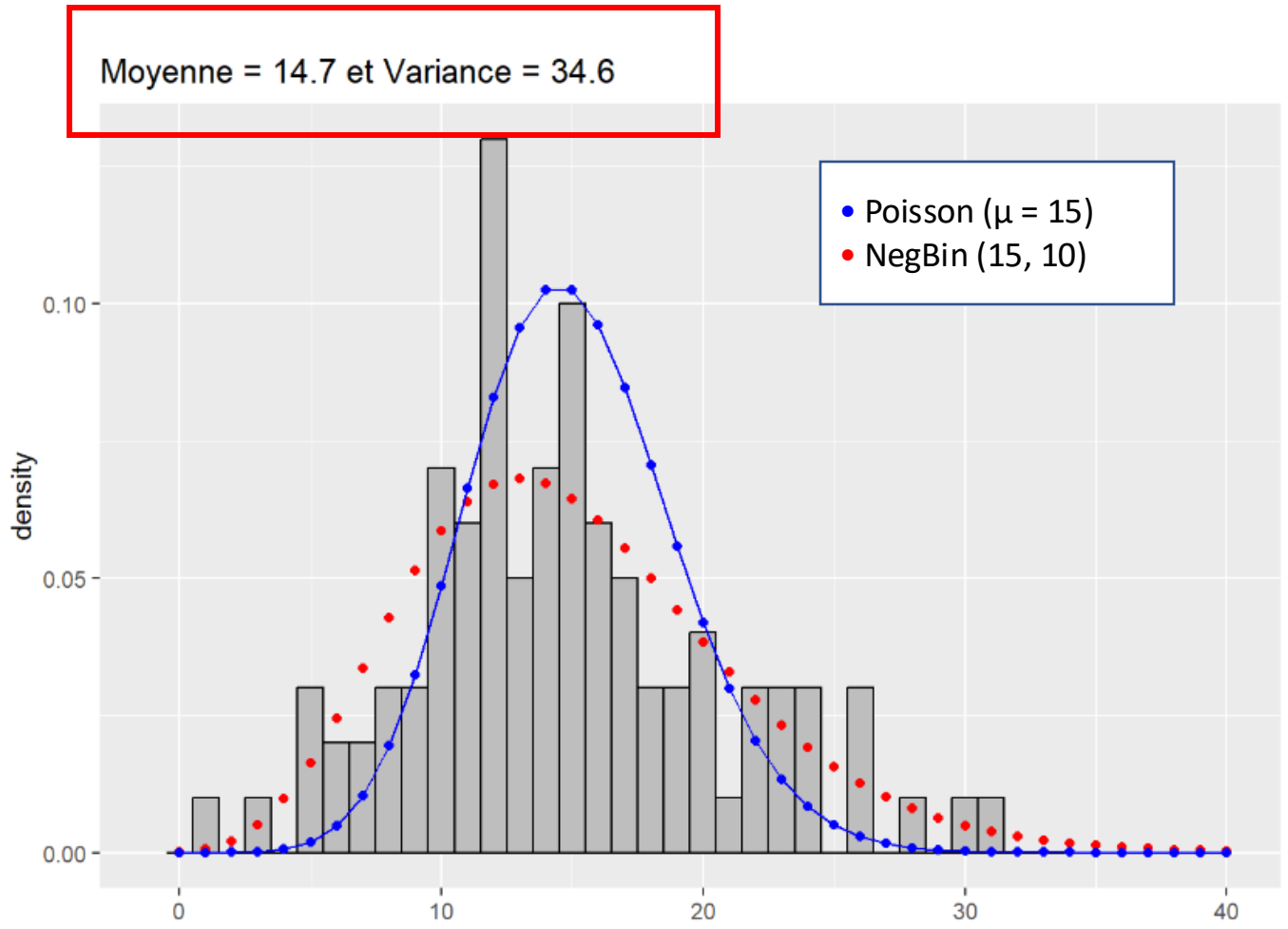
CHU Poitiers

Limite – Surdispersion (1)

- Surdispersion quand

variance$_{obs}$ > variance$_{théo}$ = espérance μ

- Dans ce cas régression de Poisson pas adaptée



Limite – Surdispersion (2)

- Quantifier la surdispersion :
  - Modèle Poisson
  - Calcul du paramètre de dispersion φ
  - $\hat{\varphi} = \frac{déviance\ résiduelle}{degrés\ de\ liberté}$
  - Si $\hat{\varphi}$ >> 1 alors surdispersion

- Régression de Poisson pas adaptée

→ distribution binomiale négative

(*Negative Binomial*, NB)

```
## Call:
## glm(formula = VFD ~ RND + AGE + SEXE, family = poisson, data = anadf)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.7311  -1.5785  -0.3328   0.7110   2.8410
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.747803   0.078140  -9.570   <2e-16 ***
## RNDB        -0.076346   0.054784  -1.394   0.1634
## AGE          0.063737   0.001464  43.522   <2e-16 ***
## SEXEH       -0.129102   0.053790  -2.400   0.0164 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2420.75  on 382  degrees of freedom
## Residual deviance:  739.65  on 379  degrees of freedom
## AIC: 1562.8
##
## Number of Fisher Scoring iterations: 5
```

```
disp.P <- modP$deviance / modP$df.residual
disp.P
```

```
## [1] 1.951577
```

Modèles inflatés en zéro

2ème exemple : Zero-inflated Negative Binomial

Negative Binomial (NB)

- Série de tirages indépendants avec une probabilité p d'obtenir un succès
- Expérience poursuivie jusqu'à obtenir n succès
- Alors la variable X représentant le **nombre d'échecs avant l'obtention d'un nombre n de succès** suit une loi binomiale négative de paramètres (n , p)

- Pour un modèle de régression, définition plus générale :
- $X \sim NB(\mu,\theta)$ avec μ (la moyenne) et θ (paramètre de « dispersion ») de sorte que $\mathrm{Var}(X) = \mu + \frac{\mu^2}{\theta}$

- Modèle NB = GLM avec une fonction de lien log (comme régression de Poisson)

Zero-inflated Negative Binomial (ZINB)

Mélange de deux processus :

- Processus de comptage et processus de Bernoulli

$$P \sim NB(\mu, \theta) \qquad Z \sim B(\pi)$$

- Variable composite Y = (1-Z) + Z*P

- Processus de Bernoulli tient compte des **zéro surajoutés** par rapport à la distribution de comptage initiale

ZINB model – mise en œuvre (1)

Variable composite Y = (1-Z) + Z*P

Analysée avec un mélange de deux modèles :

- Régression binomiale négative et régression logistique

$$log(\mu) = \beta_0 + \sum_{j=1}^{J} \beta_j X_{ij} \qquad \text{(ZINB1)}$$

$$ln(\frac{\pi}{1-\pi}) = \gamma_0 + \sum_{j=1}^{J} \gamma_j X_{ij} \qquad \text{(ZINB2)}$$

- Avec R : fonction **`zeroinfl`** du package **`pscl`**

ZINB model – mise en œuvre (2)

Zero-inflated Negative Binomial

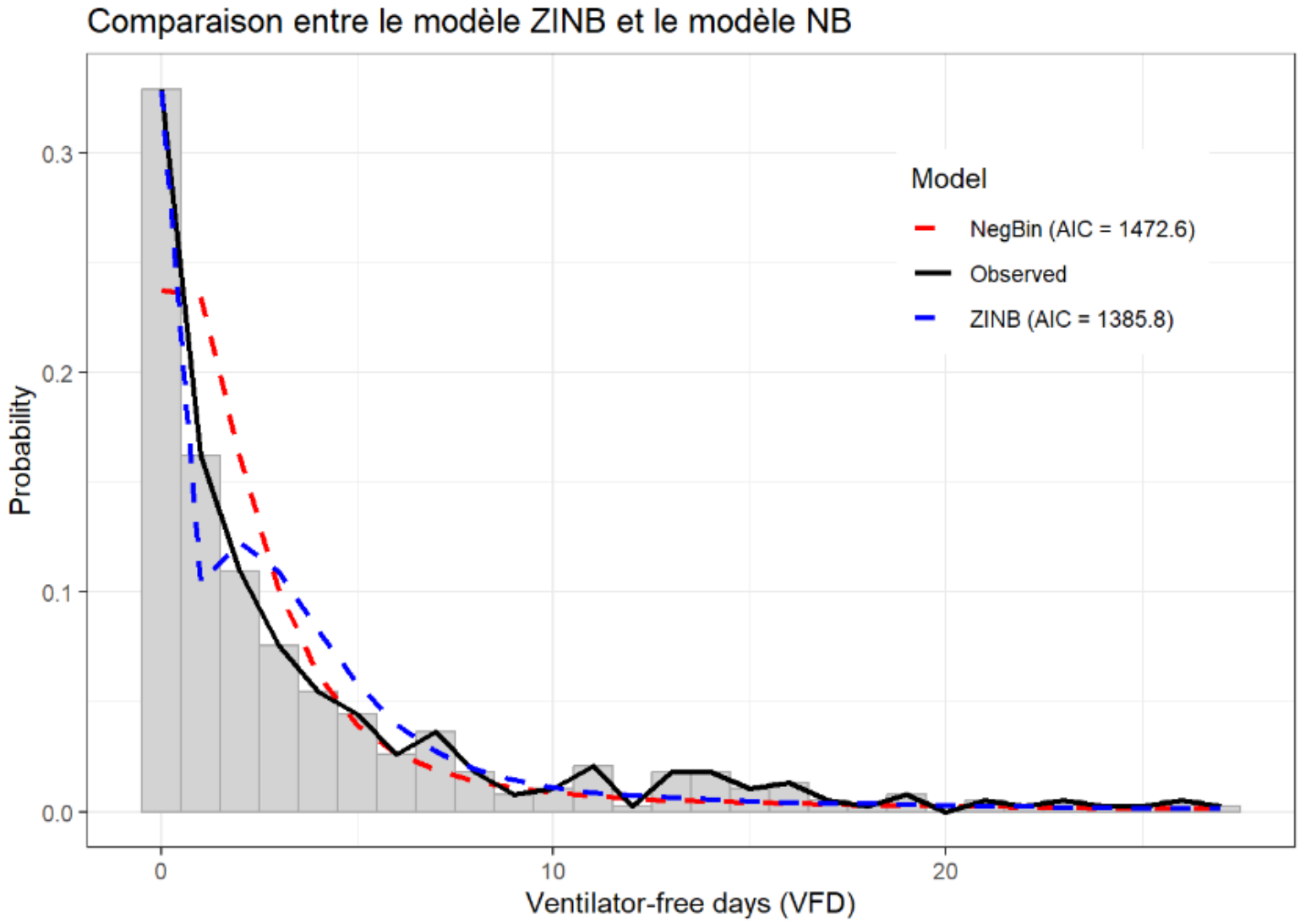
```
##
## Call:
## zeroinfl(formula = VFD ~ Age + Sex + RAND | Age + Sex + RAND, data = anadf,
##     dist = "negbin")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -2.2186 -0.6325 -0.2574  0.5069  2.6815
##
## Count model coefficients (negbin with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.426604   0.115759  -3.685 0.000228 ***
## Age          0.058390   0.002642  22.102  < 2e-16 ***
## SexMale     -0.119794   0.065787  -1.821 0.068616 .
## RANDgroupB   0.077401   0.065082   1.189 0.234329
## Log(theta)   3.319781   0.543437   6.109    1e-09 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.20052    0.77308   2.846  0.00442 **
## Age         -0.18776    0.03203  -5.862 4.57e-09 ***
## SexMale      0.30213    0.33935   0.890  0.37329
## RANDgroupB   1.75906    0.42753   4.114 3.88e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 27.6543
## Number of iterations in BFGS optimization: 15
## Log-likelihood: -683.9 on 9 Df
```

Negative Binomial

```
##
## Call:
## glm.nb(formula = VFD ~ Age + Sex + RAND, data = anadf, init.theta = 4.221840669,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4239  -1.2714  -0.2575   0.5711   1.9067
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.203058   0.122460  -9.824   <2e-16 ***
## Age          0.079243   0.002906  27.266   <2e-16 ***
## SexMale     -0.180395   0.085430  -2.112   0.0347 *
## RANDgroupB  -0.201337   0.084840  -2.373   0.0176 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(4.2218) family taken to be 1)
##
##     Null deviance: 1309.42  on 382  degrees of freedom
## Residual deviance:  457.72  on 379  degrees of freedom
## AIC: 1472.6
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  4.222
##           Std. Err.:  0.829
##
##  2 x log-likelihood:  -1462.555
```

ZINB model – comparaison avec NB

Le modèle ZINB permet de bien rendre compte de l'excès de zéro dans la distribution observée.

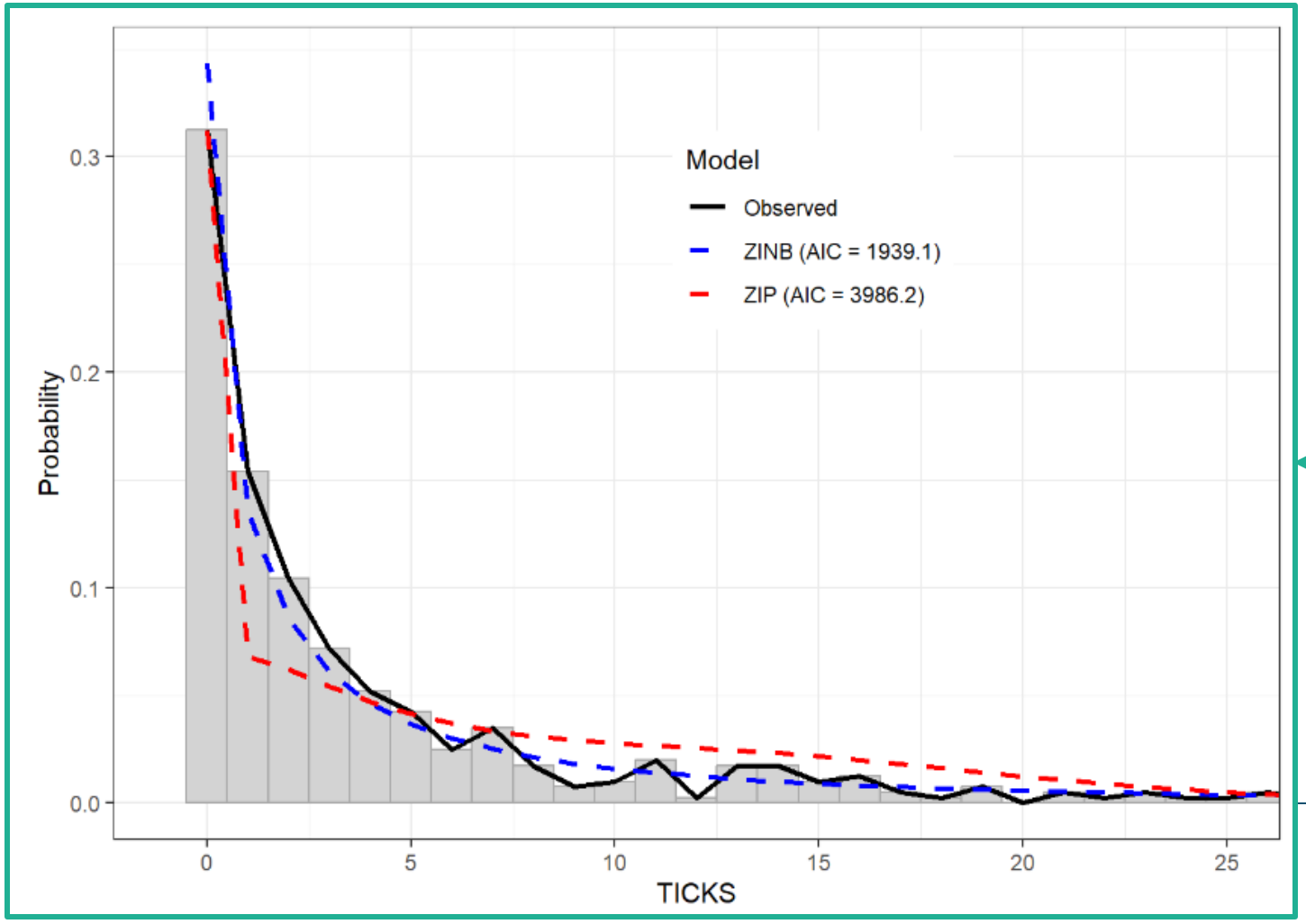


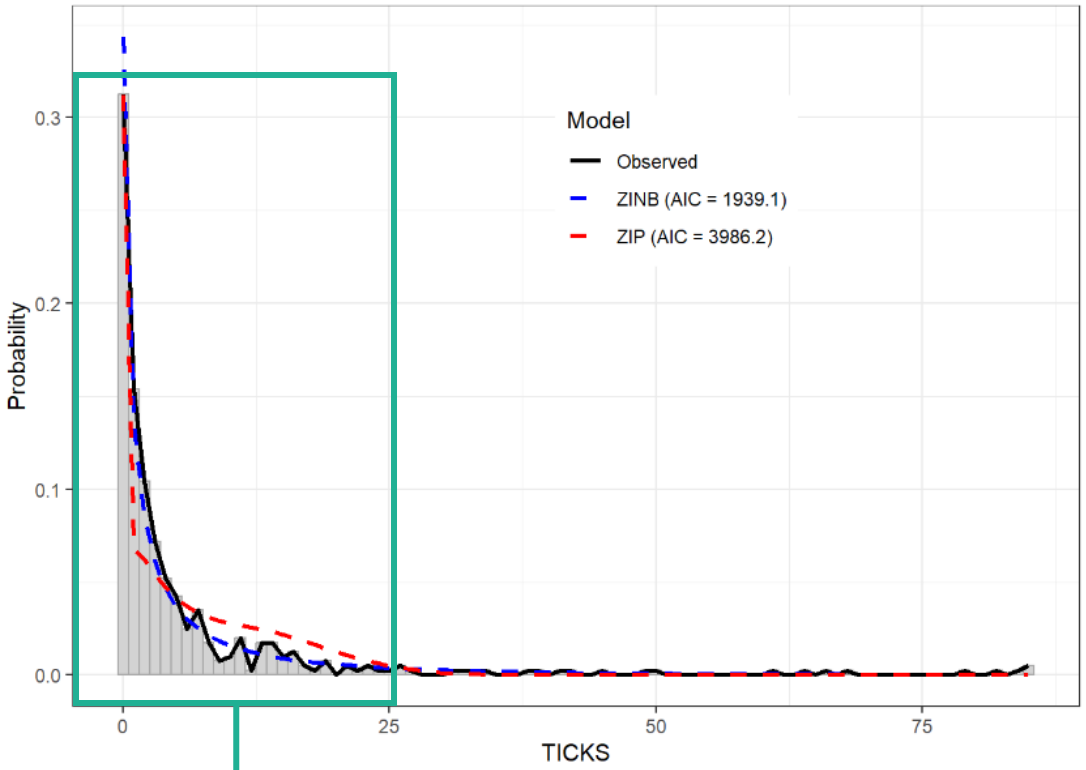
CHU Poitiers

Modèles inflatés en zéro

Lequel choisir ?

ZIP ou ZINB ? (ex1)



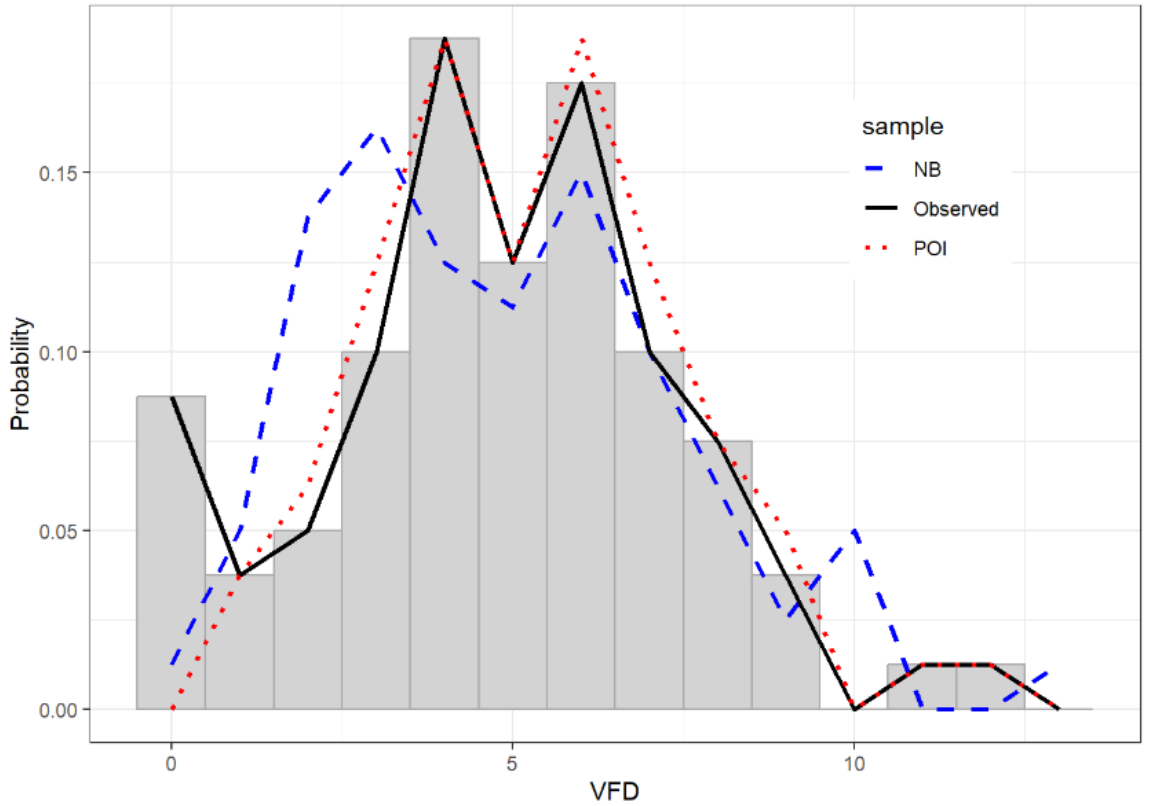


→ ZINB mieux adapté

ZIP ou ZINB ? (ex2)

- Paramètre de dispersion : $\hat{\varphi} = 1.27$ (modèle de Poisson)

```
##
## Call:
## glm(formula = VFD ~ AGE + SEXE + RND, family = poisson, data = simu)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7920  -0.4190   0.1922   0.6316   1.4717
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.147191   0.237681   0.619    0.536
## AGE          0.043357   0.006533   6.637  3.2e-11 ***
## SEXEH        0.169984   0.104458   1.627    0.104
## RNDB        -0.013433   0.105625  -0.127    0.899
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 141.186  on 79  degrees of freedom
## Residual deviance:  96.339  on 76  degrees of freedom
## AIC: 354.56
##
## Number of Fisher Scoring iterations: 5
```

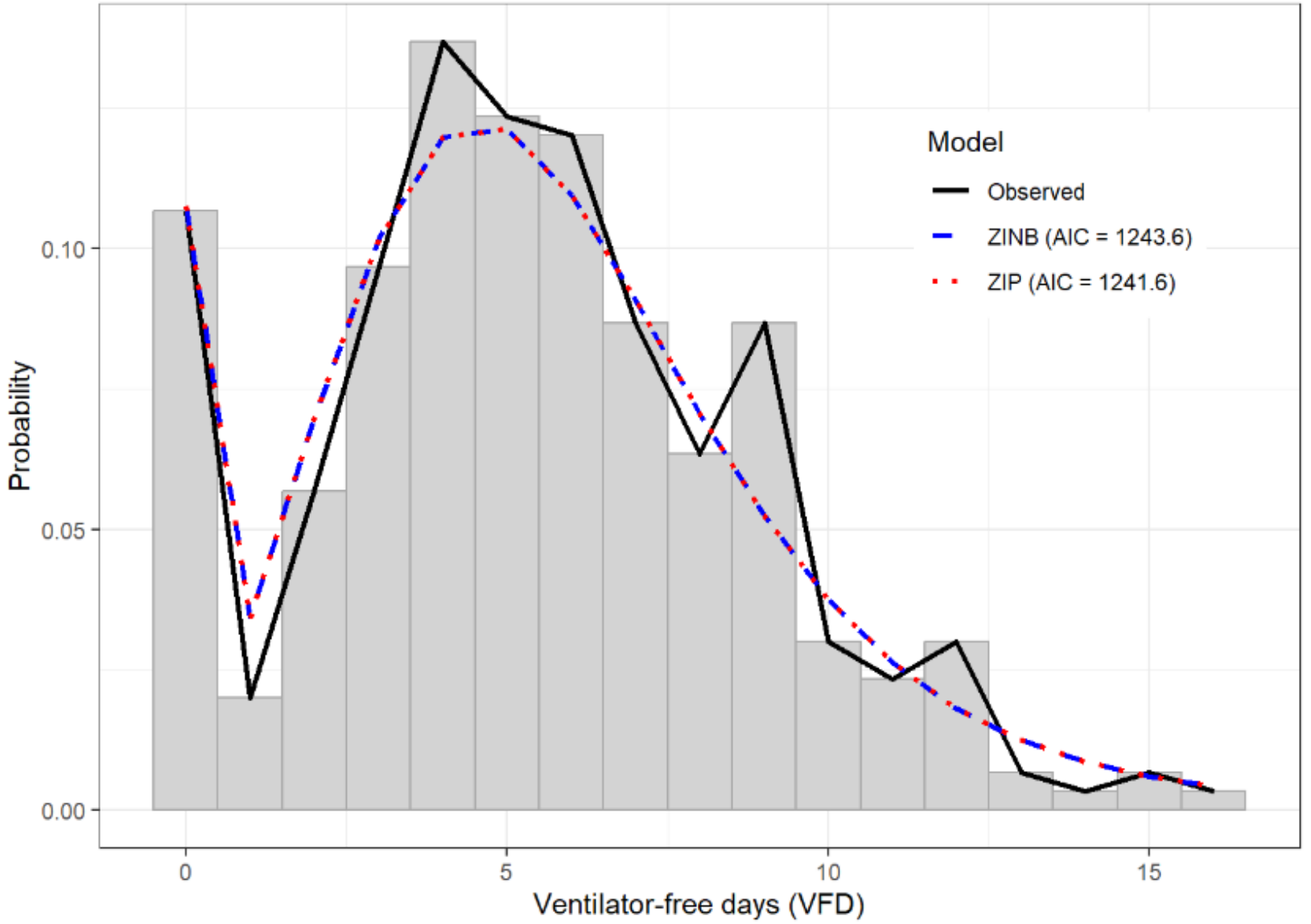


- L'échantillon simulé selon une loi de Poisson s'approche mieux du processus de comptage observé

→ ZIP mieux adapté

ZIP ou ZINB ? (ex3)

- ZINB et ZIP sont équivalents.
- On pourra privilégier ZIP (moins de paramètres)



CHU Poitiers

Merci pour votre attention

Des questions ?

Références

- Diane Lambert (1992) *Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing*, Technometrics, 34:1, 1-14

- Zeileis, A., Kleiber, C., & Jackman, S. (2008). *Regression Models for Count Data in R*. Journal of Statistical Software, 27(8), 1-25. https://doi.org/10.18637/jss.v027.i08

- Carsten F. Dormann, *Overdispersion, and how to deal with it in R and JAGS* (lien vers le pdf)

- Jeu de données **grouseticks** disponible dans le package R **lme4**, extrait de Elston, D. A., Moss, R., Boulinier ,T., Arrowsmith, C., & Lambin, X. (2001). *Analysis of Aggregation, a Worked Example: Numbers of Ticks on Red Grouse Chicks*. Parasitology 122 (05): 563-569. doi:10.1017/S0031182001007740