Loi de Poisson et régression de Poisson

Introduction
Prérequis, objectifs

- Objectifs :
  - Comprendre la Loi de Poisson
  - Comprendre la Régression de Poisson
- Prérequis :
  - Comprendre et savoir interpréter les régressions linéaires multiples
  - Avoir suivi la vidéo introductive sur les modèles log-linéaires, et leurs différentes applications

Loi de Poisson

Loi de Poisson : définition

- Mathématicien français Siméon Denis Poisson (1837)
- Soit un événement aléatoire, dont on compte le nombre d'occurrences sur une intervalle de temps fixé
- Si le nombre d'événements est en moyenne λ,
- Si la probabilité instantanée d'observer un événement est constante (pas d'effet mémoire),
- Alors la probabilité d'observer un nombre x d'événements est :

$$P(x) = \frac{e^{-\lambda}.\lambda^{x}}{x!}$$

$\lambda \in \mathbb{R}^{+}$ (réel positif)

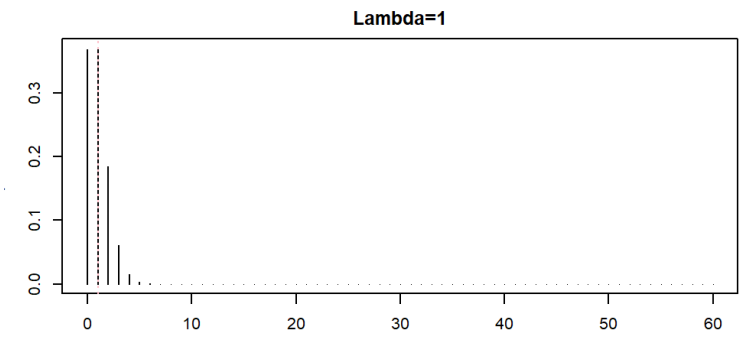
$x \in \mathbb{N}$ (entier naturel)

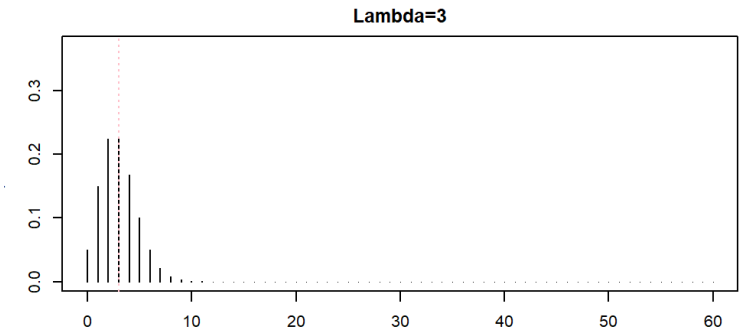
- Remarques :
  - x n'a pas de limite supérieure (mais les probabilités deviennent négligeables)
  - $\mu_x = \sigma_x^2 = \lambda$
    λ est le paramètre de cette loi, c'est aussi la moyenne et la variance

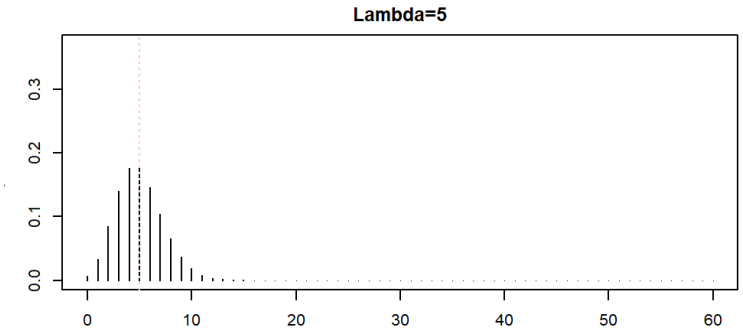
Loi de Poisson : exemples d'applications

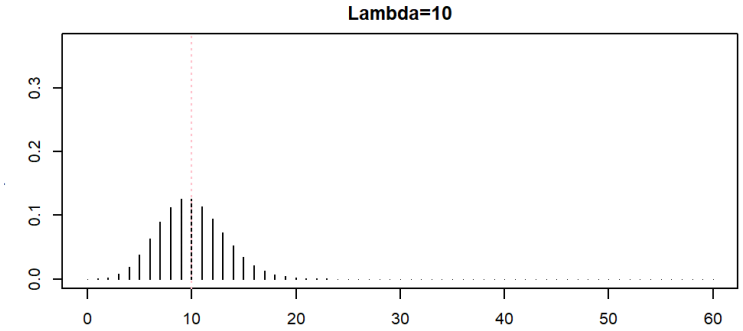
- Exemple 1 (correct) :
  - Individu statistique : un échantillon d'uranium
  - Evénement : désintégration d'un noyau d'un atome (un signal sonore sur un compteur Geiger)
- Exemple 2 (correct) :
  - Individu statistique : un service d'urgences
  - Evénement : arrivée d'un patient aux urgences
- Exemple 3 (pas forcément correct, à vérifier) :
  - Individu statistique : un patient donné
  - Evénement : survenue d'une consultation
  - Problème : la probabilité d'événement dépend des événements passés, n'est pas constante dans le temps

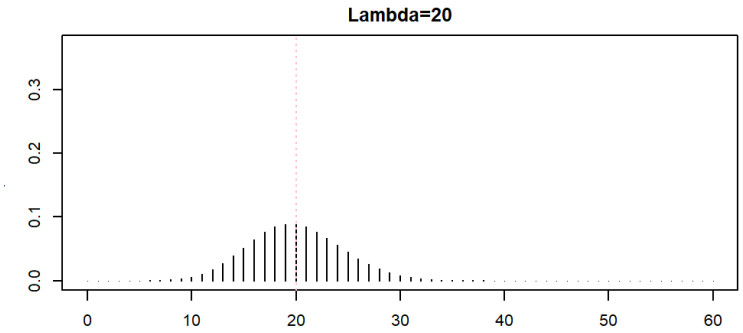
Loi de Poisson : exemples de distributions

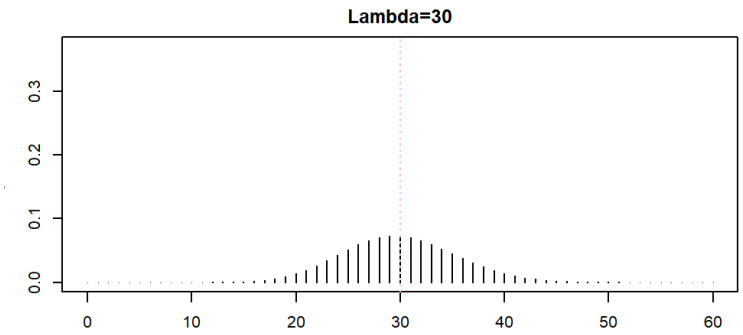






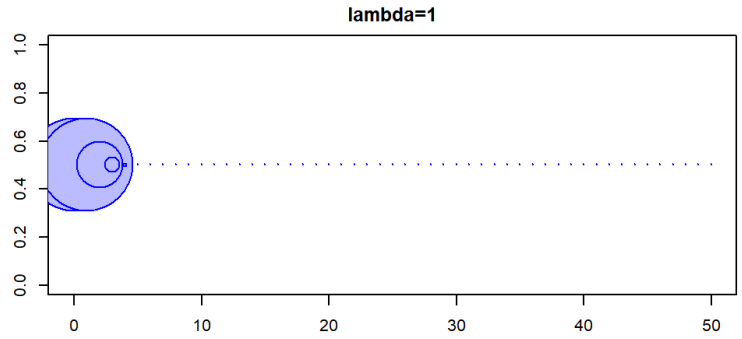


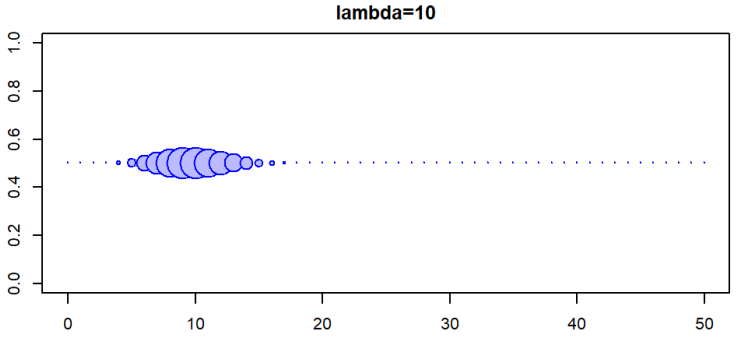


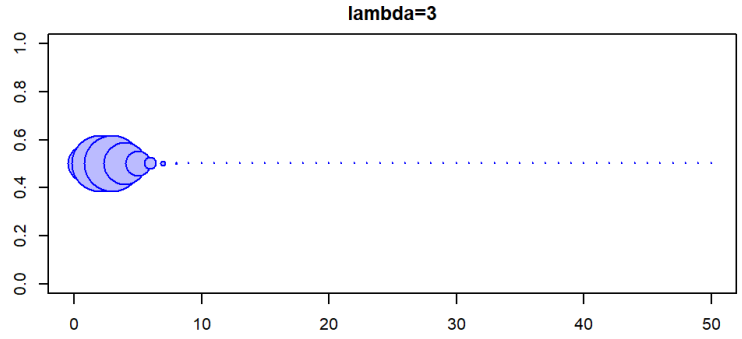


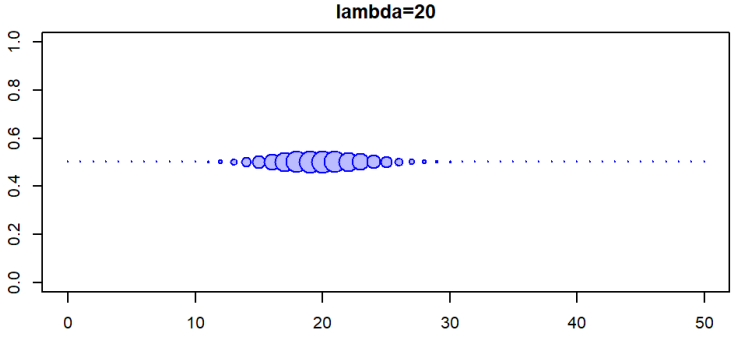
NB : λ n'est pas forcément entier

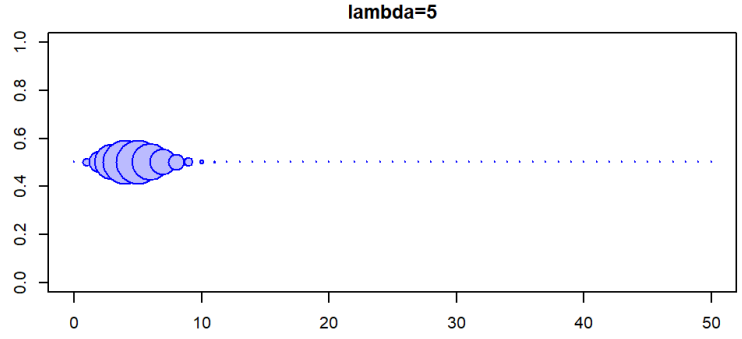
Loi de Poisson : exemples de distributions Représentation sur un seul axe

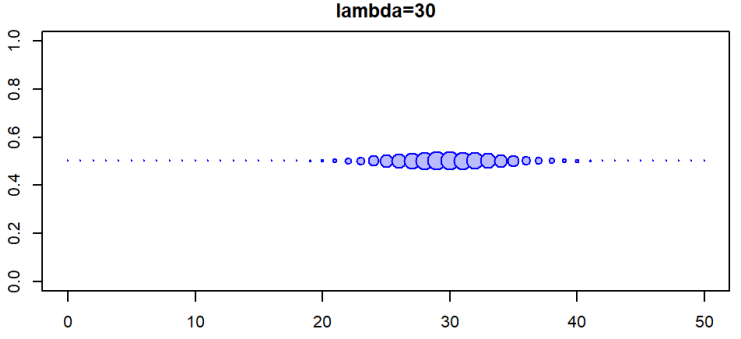












NB : λ n'est pas forcément entier

Comparaison avec la Loi Binomiale

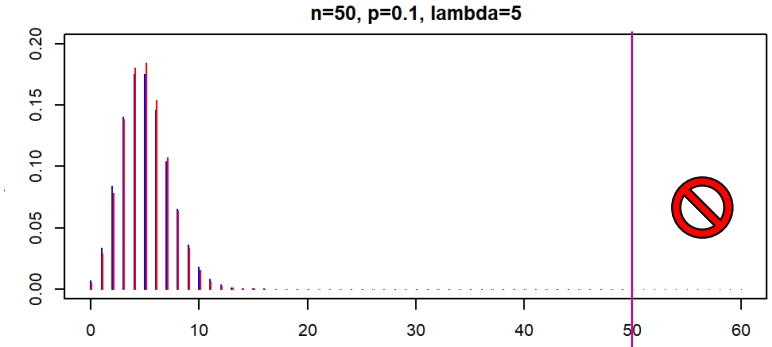
Loi de Poisson : similarité avec un processus binomial

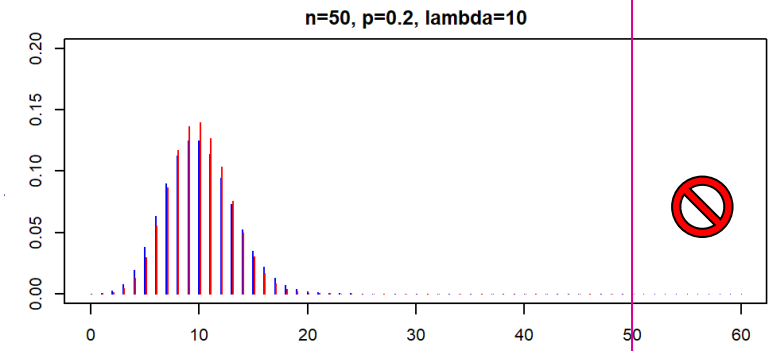
- Principe :
  - L'individu statistique initial pourrait être vu comme un ensemble d'individus statistiques (approche écologique)
  - Le nombre d'événement pourrait alors être vu comme le nombre de réalisations d'un événement binaire, pour chaque individu statistique
- Exemple 1 :
  - *échantillon d'uranium et compteur Geiger*
  - Etudier les atomes de l'échantillon
  - Statut binaire : désintégration oui/non (irréversible)
- Exemple 2 :
  - *hôpital et arrivées aux urgences*
  - Etudier les habitants d'un secteur géographique
  - Statut binaire : vient aux urgences oui/non (une seule fois)
- Exemple 3 :
  - *Nombre de consultations des patients*
  - Pas transposable

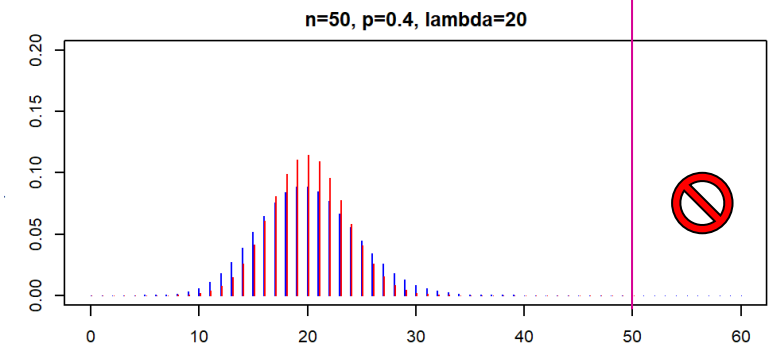
Loi de Poisson et loi binomiale

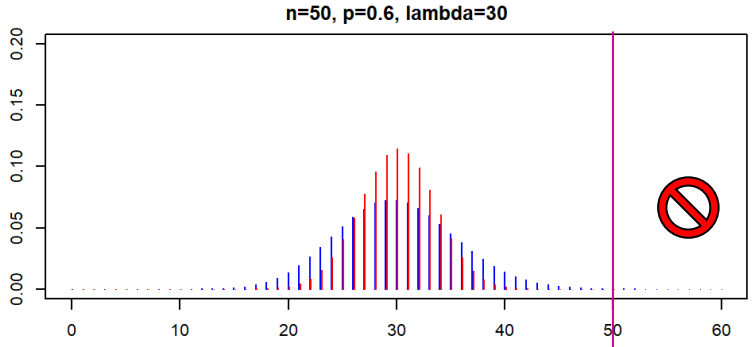
- Principe :
  - Compter le nombre d'événements binaires au sein d'une population
- Conditions :
  - Nombre d'individus quantifiable et très élevé
  - Probabilité individuelle quantifiable et très faible, évitant ainsi la saturation des proportions
  - Ne pas s'intéresser à des valeurs de x supérieures à l'effectif sous-jacent (probabilités nulles avec une binomiale, mais extrêmement faibles et non-nulles avec une Poisson)
- Similarité, si effectif et probabilités connus :
  - Effectif n, probabilité individuelle p, alors $\lambda = n.p$
  - $x \in \{0, 1, \dots, n\}$
  - Loi binomiale : $P(X = x) = p^x.(1-p)^{(n-x)}.C_n^x$
  - Loi de Poisson $P(X = x) = \frac{e^{-\mathrm{n.p}}.(\mathrm{n.p})^x}{x!}$
- NB : si effectif et probabilité non-quantifiables, mais produit λ connu, on peut utiliser Poisson mais pas binomiale. C'est un de ses intérêts.

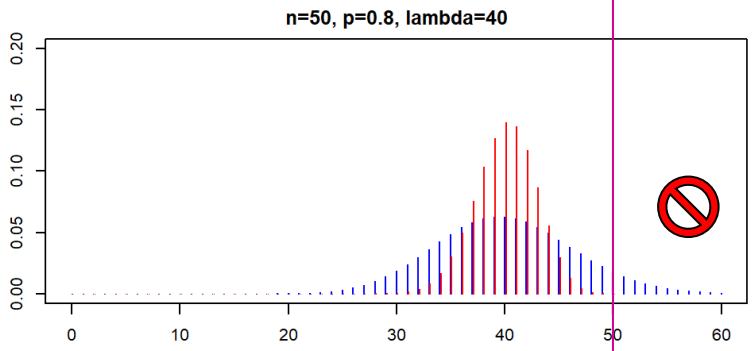
Loi de Poisson et loi binomiale

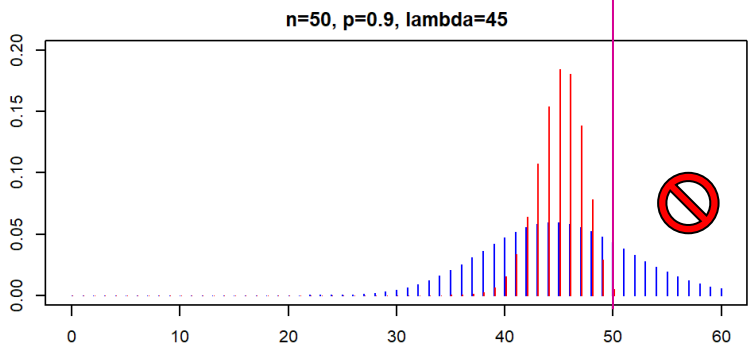










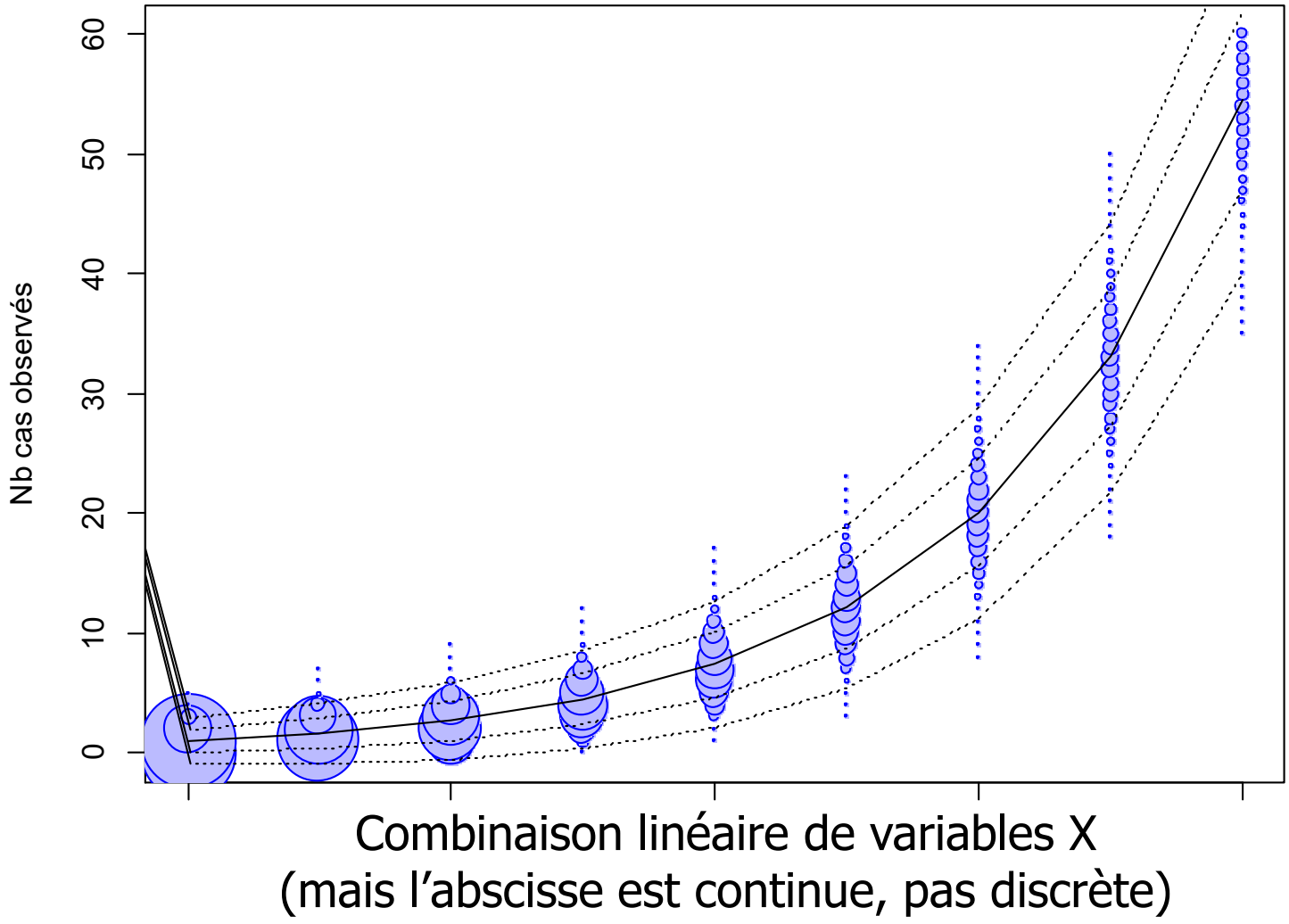


Régression de Poisson (ou modèle de Poisson)

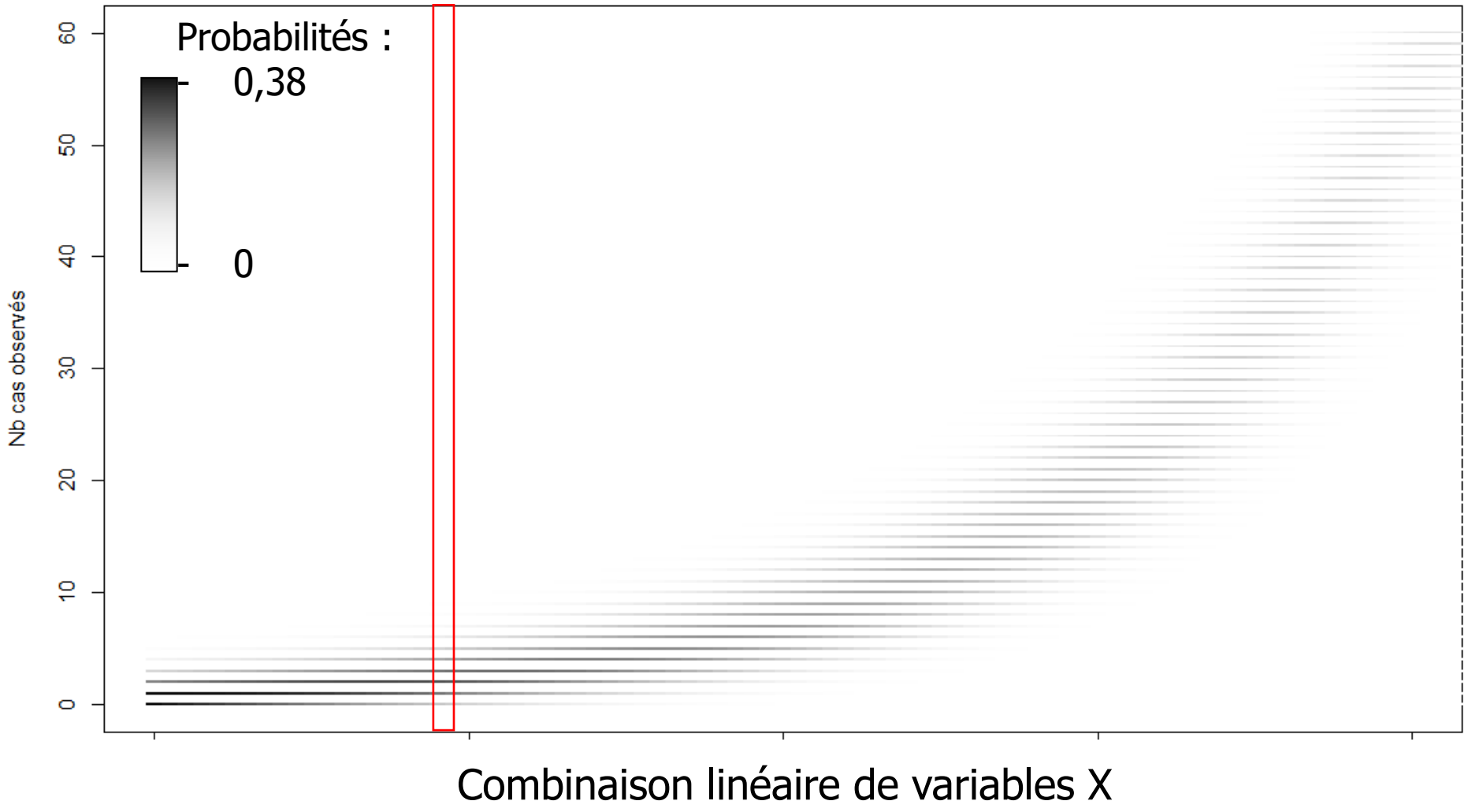
Contexte

- Variable à expliquer :
  - Y observé :
    - Nombre d'événements, entier naturel {0, 1, 2, 3…}
    - Doit suivre une loi de Poisson : conditions de validité assez précises (cf. cours dédié)
  - Y prédit :
    - Nombre moyen d'événements attendu
    - = paramètre lambda d'une loi de Poisson
- Variables explicatives :
  - Binaires, quantitatives, qualitatives (idem précédemment)
- Modèle :
  - $\ln(Y) = b_0 + b_1.x_1 + \cdots + b_i.x_i$
  - $Y = e^{(b_0 + b_1.x_1 + \cdots + b_i.x_i)}$
  - Ajustement par le « maximum de vraisemblance » pour prédire au mieux la moyenne de Y (qui est donc le paramètre λ)

Principe de la modélisation d'une loi de Poisson



Principe de la modélisation d'une loi de Poisson



Parfois besoin de prendre en compte un « effet volume »

- Avec un offset : variable pour laquelle on ne calculera pas de coefficient
  - $ln\left(\frac{Y}{offset}\right) = b_0 + b_1.X_1 + \cdots + b_i.X_i$
  - $Y = e^{(b_0 + b_1.X_1 + \cdots + b_i.X_i + \ln(offset))}$

- Exemple 1 :
  - Individu statistique : départements
  - Evénement : nombre de cas de cancers
  - Offset : nombre d'habitants de chaque département

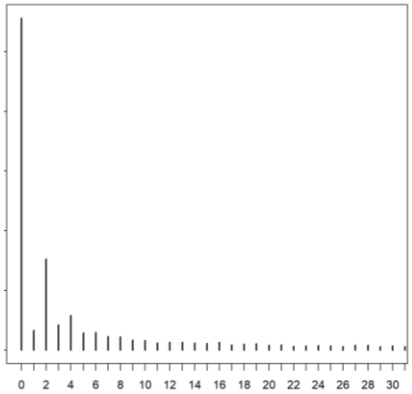
- Exemple 2 :
  - Individu statistique : patients
  - Evénement : nombre de consultations durant le suivi
  - Offset : durée du suivi en jours

- Commentaires :
  - Aucune contrainte sur l'unité ou l'échelle de l'offset, seul le ratio importera (exponentielle : sans effet mémoire)
  - Pas capable de prendre en compte une saturation : idéal pour les événements rares

Exemple simple

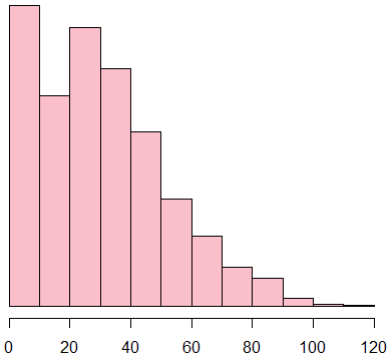
Exemple simple

- Individus statistiques : 14000 patients suivis à l'hôpital



- Variable à expliquer Y
  - nombre de consultations durant une année

- Variables explicatives X :
  - Sexe masculin (binaire, 46,3%)
  - Affection de longue durée (binaire, 6,3%)
  - Age (quantitatif, ici discrétisé)
  - Affiliation au régime général (binaire, 77,8%)



Résultat

- Exponentielles des coefficients = risques relatifs
  = facteurs multiplicateurs du nombre moyen d'événements attendu

Variable		N	Estimate		p
sexe	F	6255		Reference	
	M	5393		0.85 (0.84, 0.86)	<0.001
ALD		11648		4.85 (4.80, 4.90)	<0.001
age_d	[20,60)	6450		Reference	
	[0,20)	4146		1.02 (1.01, 1.03)	<0.001
	[60,80)	803		0.90 (0.89, 0.92)	<0.001
	[80,120)	249		0.48 (0.46, 0.49)	<0.001
regime_general		11648		1.43 (1.40, 1.46)	<0.001

Conditions de validité

Conditions de Validité

- Caractéristiques de Y, a priori
  - Y est un entier naturel
  - Attention : s'assurer que les individus avec $y = 0$ sont bien présents dans le tableau de données !
  - On ne devrait pas observer d'effet de saturation (car loi exponentielle, sans effet mémoire)
- Validation du modèle, a posteriori
  - Conditionnellement aux X, Y doit suivre une loi de Poisson
    - En particulier, conditionnellement aux X, variance=moyenne
    - Vérification : graphique des résidus
  - Pouvoir prédictif correct : R^2
  - Parcimonie des variables explicatives
  - Absence de multicolinéarité
  - Absence d'individus trop influents

Analyse des résidus standardisés de Pearson

- Contrairement à la régression linéaire, les résidus doivent augmenter avec la moyenne prédite car $\mu_x = \sigma_x^2 = \lambda$
- Résidus de Pearson (standardisés) :
  - $r_{p_i} = \frac{r_i}{\sqrt{\widehat{\lambda_i}}} = \frac{y_i - \widehat{\lambda_i}}{\sqrt{\widehat{\lambda_i}}}$
  - y_i : pour l'individu i, valeur observée (0, 1, 2…)
  - $\widehat{\lambda_i}$: pour l'individu i, c'est à la fois :
    - Le nombre *moyen* d'événements prédit
    - Le paramètre de Poisson prédit
    - Le carré de l'erreur de prédiction
- Les résidus de Pearson doivent être :
  - De distribution normale (histogramme)
    - Symétrique par rapport à 0
    - 95% des valeurs entre -1,96 et +1,96
  - Indépendants de la valeur prédite λ (nuage de points)
    - Moyenne mobile reste nulle
    - Dispersion mobile reste constante

Diagnostic de surdispersion ou sousdispersion

- Peut être fait directement sur la distribution des résidus de Pearson
- Ou plus simplement, calculer :
  - $\frac{Somme\ des\ résidus\ de\ Pearson}{Nombre\ de\ degrés\ de\ liberté} = \frac{\sum r_{p_i}}{n-p-1}$
  - Devrait être proche de 1
  - « surdispersion » si nettement supérieur
  - « sousdispersion » si nettement inférieur
  - Il existe un test, mais…

Et lorsque la régression de Poisson n’est pas valide ?

Autres modèles

- Problème principal du modèle de Poisson :
  - Hypothèse très forte : le décompte suit une loi de Poisson
  - Conditionnellement aux X, moyenne=variance=lambda. Souvent invalide :
    - « sur-dispersion » si variance > moyenne
    - « sous-dispersion » si variance < moyenne
- Autres modèles si hypothèse invalide :
  - Quasi-Poisson :
    - modélise la variance comme une fonction linéaire de la moyenne
    - Interprétation identique
  - Loi binomiale négative
    - modélise la variance comme une fonction linéaire de la moyenne et de son carré
    - Interprétation identique
  - Modèle de Poisson « zero-inflated »
    - Première étape de régression logistique : y=0 ou y>0 ?
    - Deuxième étape de régression de Poisson, que si y>0 (mais peut prédire aussi y=0…)
    - Donc deux jeux de coefficients : des OR pour la première étape, des RR pour la deuxième
  - Modèles de sauts (hurdle)
    - Principe similaire, mais pas de risque de prédire y>0 puis y=0
  - Etc.

Merci de votre attention !
RDV sur
http://objectifthese.org

http://objectifthese.org
contact@objectifthese.org

Objectif Thèse