AUTRES LOIS, AUTRES MODÈLES : Introduction

- Loi de Poisson
- Loi Binomiale Négative
- Loi Exponentielle et loi Gamma

Plan commun pour chaque loi étudiée

Exemple

Présentation

Hypothèses

Variance

Le modèle dans le cadre GLM

Les écarts aux hypothèses initiales

Quasi-vraisemblance, surdispersion

Les autres lois étudiées parmi celles les plus rencontrées dans la pratique sont les lois de Poisson, Binomiale Négative, Exponentielle, Gamma.

Bibliographie :

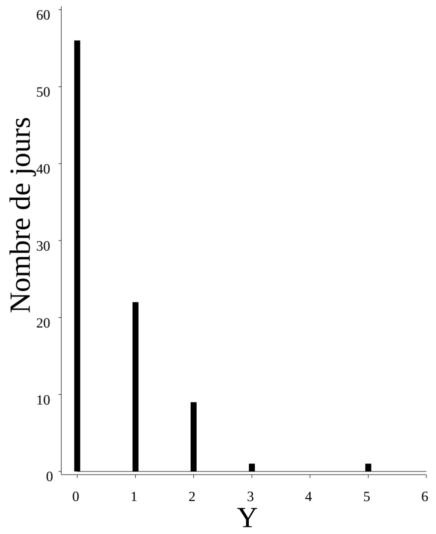
- *Biostatistique*, Bruno Scherrer, Gaëtan Morin éditeur (1984).
- *Univariate discrete distribution*, Johnson & Kotz, Wiley (1993).
- *Continuous univariate distribution*, Johnson & Kotz, Wiley (1993).
- *Cours de calcul des probabilités*, Gérard Calot, Dunod Décision (1967).
- *Probabilités, analyse des données et statistique*, Gilbert Saporta, éditions Technip (1990).
- *Analysis of infectious disease data*, N.G. Becker, Chapman & Hall (1989).
- *Models in biology*
  Mathematics, Statistics and Computing, D. Brown, P. Rothery, Wiley (1993).
- *Modèle de prédiction de l'occurrence des mammites cliniques chez la vache laitière au cours de ses lactations successives*, P. Gasqui & O. Pons, à paraître.
- *Initiation aux processus aléatoires*, S. Carlin, DUNOD.

LOI DE POISSON : exemple

Exemple : le merle à plastron

Y	0	1	2	3	4	5	6
Fréquence	56	22	9	1	0	1	0

Y : nombre de merles capturés par jour



Exemple : "*le merle à plastron (*Turdus torquatus*) est un oiseau qui en automne erre dans les bois clairs et les buissons des montagnes, entre 1500 mètres et 2000 mètres d'altitude.*

En 1968, à la station ornithologique du Col de Golèze située dans les Alpes françaises à 1700 mètres d'altitude, 48 merles à plastron ont été capturés au filet japonais durant les 89 jours d'ouverture de la station." (Bruno Scherrer, Biostatistique).

Soit Y le nombre de merles capturés par jour, à l'aide de 4 pièges. Si la répartition dans le temps (et dans l'espace) des captures de merles à plastron est aléatoire, et si la probabilité de capture reste constante, la distribution de probabilité de Y est une loi de Poisson. En effet, chaque événement consiste en la capture d'un merle. La probabilité instantanée (dans un petit intervalle de temps) de capture est très faible, le nombre de captures par jour est petit, il varie entre 0 et 5. Il est entier, il n'est pas borné, si ce n'est par la limite du nombre de merles qu'il est matériellement possible de capturer en un jour avec 4 pièges; rien n'empêche qu'il soit très grand, si ce n'est la probabilité d'une telle réalisation, qui est infiniment petite.

Si les filets restent à la même place au cours de la saison (ce qui est le cas), si les oiseaux ne s'habituent pas aux pièges et ne les évitent pas, si leur comportement n'est ni grégaire ni territorial ou répulsif, les événements sont indépendants et la probabilité de capture reste constante d'un moment à l'autre.

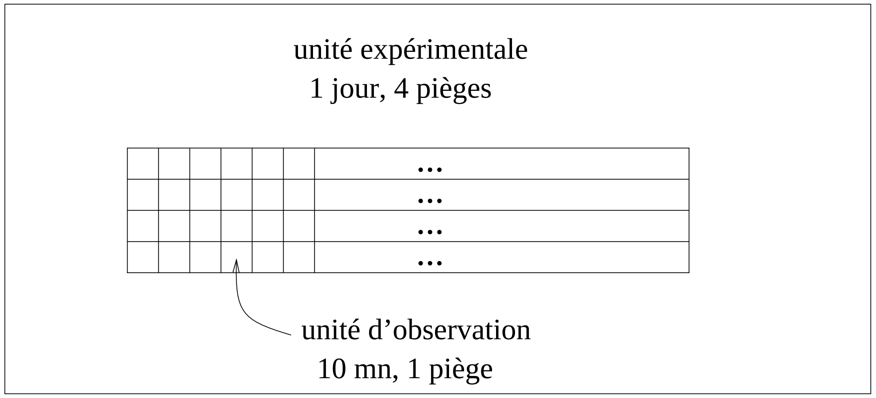
$Y \sim \mathcal{P}(np)$, où np est le nombre moyen de merles capturés par jour. $n\widehat{p} = 48/89$.

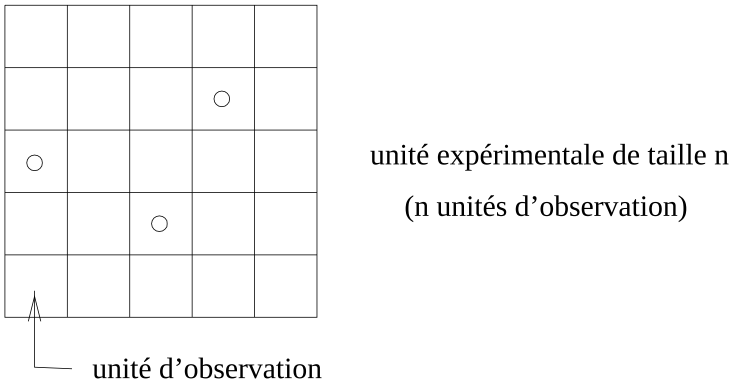
LOI DE POISSON :

unité d'observation

unité expérimentale

merle à plastron





Autres exemples classiques :

On utilisera la distribution de Poisson pour calculer par exemple la probabilité qu'un nombre donné d'appels (0, 1, 2, 3 ...) parviennent à un standard téléphonique en une certaine période de temps ou encore pour évaluer la probabilité d'observer un certain nombre de défauts par unité de surface d'une plaque métallique servant à la fabrication de véhicules lourds.

Le nombre d'accidents par semaine sur une autoroute, le nombre d'arrivées par heure à un guichet d'une compagnie aérienne, le nombre de voitures par minute arrivant à un poste de péage d'une autoroute, le nombre de pots de café vendus en une journée dans un supermarché, le nombre de naissances par année dans une petite municipalité, le nombre d'insectes venant se poser sur une fleur durant une période de temps, le nombre d'animaux d'un troupeau atteints par une certaine pathologie, le nombre d'événements pathologiques touchant un animal durant une année sont autant de variables qui pourront être analysées à l'aide de la distribution de Poisson.

D'autres exemples font intervenir la notion d'espace, de région, de volume. Nous pouvons penser, par exemple, au nombre d'erreurs typographiques dans une page de journal, au nombre de nids de poule sur un kilomètre de route, au nombre d'arbres d'une certaine espèce végétale dans un certain territoire ou encore au nombre de grains de pollen dans un mètre cube d'air, et au nombre de poissons d'une espèce déterminée et relativement rare, pêchés avec un certain type de filet.

L'observation poissonienne est mesurée sur une unité expérimentale. L'unité expérimentale peut être découpée en unités d'observations. Ce découpage est plus ou moins naturel suivant le contexte. Cela peut être le découpage d'un intervalle de temps, d'une région de l'espace, ou d'un ensemble d'objets, selon la nature de l'unité expérimentale.

Dans le cas du merle à plastron, l'unité expérimentale est l'ensemble des 4 pièges observés pendant une journée. L'unité d'observation est LE piège observé pendant un "petit intervalle de temps", que l'on peut arbitrairement définir comme un intervalle de 10 minutes, 10 minutes étant à peu près le temps nécessaire pour relever un piège et le remettre en service. L'unité d'observation est suffisamment petite pour qu'il ne soit pas possible d'observer l'occurrence du phénomène plus d'une fois sur une unité d'observation. La probabilité d'occurrence de l'évènement est très faible, pour une unité d'observation.

Remarque : le fait que la probabilité reste constante d'une unité d'observation à l'autre assure que le nombre d'événements pendant une période de temps donnée ne dépende que de la durée de la période de temps, condition nécessaire, dans la définition de la loi de Poisson.

LOI DE POISSON : présentation

C'est la loi d'une variable aléatoire Y

qui exprime le nombre d'occurrences

d'événements

distribués au hasard

dans un certain laps de temps

ou dans une région donnée de l'espace.

Le type de l'événement doit être bien identifié.

Le "laps de temps" ou la "région de l'espace" fait référence à la notion d'unité expérimentale.

L'événement est observé sur une unité d'observation.

L'unité expérimentale est constituée d'un ensemble de n unités d'observation ($Y \leq n$).

L'unité d'observation peut être une durée unitaire, un segment unitaire, une surface ou un volume unitaire permettant la réalisation ou l'observation d'au plus un événement.

L'unité expérimentale est alors une durée totale, un grand segment, une grande surface ou un grand volume constitué de n unités d'observation.

Pour un ensemble d'unités expérimentales représentatives de la population étudiée, et sous certaines hypothèses, Y suit une distribution de Poisson.

La loi de Poisson est une distribution discrète qui dérive de la loi Binomiale. Elle correspond à une distribution Binomiale où l'une des éventualités a une probabilité très faible.

LOI DE POISSON : hypothèses

sur les unités expérimentales

- n, taille des unités expérimentales fixée a priori et *constante*.

sur le phénomène observé

- *événement* compté *bien défini*, *identique* pour toutes les unités expérimentales.
- *comptage exhaustif* par unité expérimentale.
- *un* événement compté *au plus* par unité d'observation.
- *indépendance* des *unités d'observation*.
- *répartition* des événements *au hasard* parmi les unités d'observation d'une unité expérimentale.
- phénomène "*rare*".

Sur les unités expérimentales :

- L'unité expérimentale étudiée ou observée doit avoir une durée, une longueur, une surface ou un volume qui doit avoir été fixé *a priori*, et qui doit être identique pour toutes les unités expérimentales.

Sur le phénomène observé :

- Le type de l'élément ou de l'événement compté sur chaque unité expérimentale doit être bien défini et identifiable. Ce comptage doit être exhaustif sur chaque unité expérimentale. Donc toutes les unités d'observation doivent être regardées. On ne peut avoir plus d'un événement observé sur une unité d'observation.
- Les unités d'observation doivent être indépendantes les unes des autres.
- Les événements observés doivent se répartir au hasard sur l'ensemble de l'unité expérimentale observée. La fréquence des événements est constante sur toute la durée d'observation ou sur l'ensemble de la région observée. (la répartition des éléments est homogène sur l'unité expérimentale).

  Dans ce cas la probabilité d'observer un événement sur une unité d'observation, dans un laps de temps très court, sur une petite longueur, sur une petite surface ou dans un petit volume est constante (si les conditions sont constantes, c'est à dire si les niveaux des facteurs et les valeurs prises par les variables explicatives sont constants sur l'unité d'observation).
- Le phénomène doit être "rare" : le nombre espéré $\mathrm{E}(Y)$ d'événements observés doit être "petit" par rapport au nombre d'unités d'observation n de l'unité expérimentale :
  $\mathrm{E}(Y)/n < 0.10$ ou < 0.05 en général. La probabilité d'observer un événement sur une unité d'observation est faible.

LOI DE POISSON : variance

Distribution discrète.

Pour une unité expérimentale de taille n

$$Y \sim \text{Poisson}(np)$$

$$\boxed{\text{Var}(Y) = \text{E}(Y)}$$

np : effectif moyen.

n : taille des unités expérimentales

p : probabilité constante d'occurrence de l'événement
par unité d'observation.

Si l'on convient que la dimension de l'unité d'observation vaut 1, le nombre d'unités d'observation dans une unité expérimentale n est aussi la dimension ou taille de l'unité expérimentale.

La taille n est la durée totale T, la longueur totale L, la surface totale S ou le volume total V de l'unité expérimentale.

Le paramètre p s'interprète comme la probabilité d'occurrence de l'événement par unité d'observation, l'unité d'observation pouvant être une unité de temps, de longueur, de surface, de volume, ...

LOI DE POISSON :

comme approximation

de la loi Binomiale

La loi $\mathcal{B}(n,p)$ tend vers une loi $(\mathcal{P}(np))$

quand

- n tend vers l'infini (n grand)
- et p tend vers 0 (événement "rare").

Le nombre Y d'événements observés parmi n unités d'observation suit une distribution Binomiale de paramètres (n,p). On a

$$P(Y=y) = \mathcal{C}_n^y p^y (1-p)^{n-y} \text{ pour } y = 0, 1, \ldots, n$$

et

$$P(Y=y) = 0 \text{ pour } y > n$$

p est la probabilité de réalisation de l'événement sur les n unités d'observation.

$\mathrm{E}(Y) = np \qquad \mathrm{Var}(Y) = np(1-p)$

Cette distribution tend vers une distribution de Poisson de paramètre np telle que :

$$P(Y=y) = \frac{(np)^y}{y!} \exp(-np) \text{ pour } y \geq 0$$

La loi de Poisson est la loi des événements rares.

Remarque : on peut présenter la loi de Poisson dans le cas des merles à plastron comme la limite d'une loi Binomiale $\mathcal{B}(n,p)$, où n la taille de l'unité expérimentale est grand, et p la probabilité de capture par unité d'observation, c'est à dire par petit intervalle de temps, tend vers 0.

LOI DE POISSON :

le modèle dans le cadre GLM

Y : nombre d'événements
sur une unité expérimentale,

$$Y \sim \mathcal{P}$$

lien canonique : lien log,

$$\boxed{\mathrm{Var}(Y) = \mathrm{E}(Y)}$$

Vérification empirique :

dans le cas où l'on dispose de répétitions par combinaison des modalités des facteurs étudiés, il suffit de tracer le graphique des moyennes en fonction des variances estimées. On vérifie que les points varient aléatoirement autour de la droite de pente 1.

LOI DE POISSON :

les écarts aux hypothèses initiales

Taille des unités expérimentales non constante

- Modélisation de la loi de n
- Introduire n sous forme d'un "offset"

Unités d'observation non indépendantes

- Introduire un facteur de dispersion
- Modélisation de la dépendance

Unités expérimentales non indépendantes

- Introduire un facteur de dispersion
- Prise en compte de la dépendance (GEE, ...)

Taille des unités expérimentales non constante,

c'est à dire nombre d'unités d'observation non constant.

Le "laps de temps" ou la "région de l'espace" qui définit l'unité expérimentale n'a pas la même taille pour toute les unités expérimentales. Le paramètre np de la loi de Poisson apparaît différent car n n'est pas identique pour toutes les unités expérimentales.

Le nombre de pièges est par exemple différent d'un jour à l'autre.

Solutions possibles

Si on connait la loi de n, on l'introduit dans le modèle et on adapte la loi de Poisson initiale à cette connaissance (par exemple, la Binomiale Négative est une adaptation de la loi de Poisson, au cas où n est aléatoire et suit une loi Gamma). Sinon on peut introduire n sous forme d'un "offset" dans le modèle GLM (exemple de la modélisation du nombre de suicides par mois, avec un lien log, $\log(\frac{\text{nb suicides}}{\text{nb jours}}) = \log(\text{nb suicides}) - \log(\text{nbjours})$, $\log(\text{nb jours})$ est introduit en offset).

Non indépendance des unités d'observation

C'est le cas si un animal piégé en attire d'autres, ou bien les éloigne, ou si en général l'attitude des autres animaux est modifiée. C'est aussi le cas si la capture d'un animal modifie l'efficacité ultérieure du piège, ou bien si les pièges sont trop rapprochés les uns des autres et les déplacements des merles non indépendants les uns des autres.

Solutions possibles

Introduire un facteur de dispersion dans le modèle (en utilisant par exemple une loi Binomiale Négative, voir p. 17 et 18). On peut aussi introduire une modélisation de cette dépendance dans le modèle en passant à la modélisation des intervalles de temps entre captures; la probabilité de capture est modifiée ou pas après une capture précédente.

Non indépendance des unités expérimentales

C'est le cas quand les captures effectuées un jour donné modifient la capacité de capture du ou des jours suivants.

Solutions possibles

Modéliser la matrice de corrélation entre unités expérimentales en utilisant par exemple la méthode GEE (generalized estimating equation) qui est une extension des modèles GLM permettant de prendre en compte une structure de corrélation.

LOI DE POISSON :

les écarts aux hypothèses initiales (suite)

Non répartition des événements
au hasard
ou
fréquence non constante

- Redéfinir l'unité expérimentale
- Introduire un facteur de dispersion
- Utiliser un facteur aléatoire
- Prendre en compte les facteurs de variation

Non répartition des événements au hasard ou fréquence non constante

Oubli de facteurs de variation dans le “plan d'expérience” qui fait que la probabilité de capture
(1) varie au sein d'une unité expérimentale : elle peut être différente d'un piège à l'autre, si les pièges sont dans des zones différentes vis à vis de l'habitat de l'animal et de sa densité; elle peut être différente entre le jour et la nuit
(2) varie d'une unité expérimentale à l'autre : elle peut être différente d'un jour à l'autre, si l'activité des animaux varie selon les conditions climatiques, la période de l'année, ...

Solutions possibles

• Redéfinir l'unité expérimentale de façon à ce que la fréquence de l'événement soit constante, sur les unités expérimentales, dans des conditions constantes.
• Prendre en compte ces facteurs de variations dans le modèle quand ils sont connus et accessibles.
• Faire intervenir un facteur aléatoire dans le modèle qui les prendra globalement en compte.
• Introduire un facteur de dispersion dans le modèle.

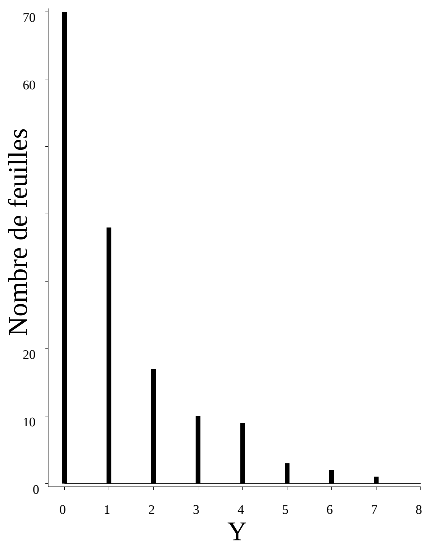
Remarque : on dit ici que la probabilité d'occurrence de l'événement doit être constante sur l'unité expérimentale, pour que l'on ait une loi de Poisson. En fait, c'est imposer une contrainte trop forte, pas nécessaire dans tous les cas. La probabilité doit être constante sur l'unité expérimentale, ou bien “constante par morceaux”, le découpage en morceaux de l'unité expérimentale étant le même pour toutes les unités expérimentales.
Imaginons que l'on s'intéresse au nombre d'accidents de la route par jour. Le jour est l'unité expérimentale. La probabilité d'accident n'est pas constante sur toute la durée du jour. Par contre, on peut supposer qu'elle est constante sur chacune des 6 plages de 4 heures d'une journee. Sur chacune des périodes de 4 heures, le nombre d'accidents suit une loi de Poisson. La somme de deux lois de Poisson étant une loi de Poisson, le nombre d'accidents par jour suit une loi de Poisson.

LOI BINOMIALE NÉGATIVE : exemple

Exemple : la tétranyque rouge du pommier

Y	0	1	2	3	4	5	6	7	8+
Nombre de feuilles	70	38	17	10	9	3	2	1	0

Y : nombre d'individus par feuilles



Exemple : la tétranyque rouge du pommier.

Bliss et Fisher ("Fitting the binomial distribution to biological data and a note on the efficient fitting of the negative binomial", Biometrics, 1953, 9:176-200), dans une étude sur la tétranyque rouge du pommier, ont cherché le type de répartition spatiale de ces acariens. Pour ce faire, ils ont compté le nombre Y d'individus sur 150 feuilles de pommier.

L'expérimentateur note des agrégations dans la répartition des individus, la distribution de fréquence du nombre de tétranyques observées par feuille obéit à une distribution Binomiale Négative.

La loi Binomiale Négative permet de prendre en compte une certaine dépendance des unités d'observation entre elles, qui fait que la distribution de Poisson ne s'ajuste pas correctement.

Dans les exemples où le taux de croissance ne dépend que d'un taux d'immigration constant, la loi de Poisson convient.

LOI BINOMIALE NÉGATIVE : présentation

Présentation classique :

C'est la loi d'une variable aléatoire Y' qui exprime

le nombre d'essais indépendants nécessaires pour obtenir k occurrences d'un événement

ayant une probabilité p constante d'être observé dans chaque essai.

Autre présentation :

C'est aussi la loi d'une variable aléatoire Y qui exprime

le nombre de non événements nécessaires pour observer k occurrences d'un événement.

Il existe deux types de présentations de la loi binomiale négative, la présentation classique, et l' "autre présentation". On verra plus tard que c'est l'autre présentation qui conduit à écrire que la variance est proportionnelle à l'espérance, comme on le suppose lorsqu'on utilise un GLM avec loi binomiale négative. La première présentation a l'avantage d'être classique, la deuxième a l'avantage d'être cohérente avec l'hypothèse de proportionnalité entre la variance et le carré de l'espérance.

On note Y' la variable associée à la présentation classique, et Y la variable associée à l'autre présentation.

La loi binomiale négative est la loi d'une variable aléatoire Y' qui est le nombre de tirages avec remise qu'il faut effectuer pour obtenir un nombre k de boules blanches dans une urne de Bernoulli composée d'une proportion p de boules blanches. C'est aussi la loi d'une variable aléatoire Y qui est le nombre de boules noires qu'il faut tirer (tirage avec remise) dans une urne de Bernoulli composée d'une proportion p de boules blanches, et $1-p$ de boules noires.

Y' est le nombre d'unités d'observation nécessaires à la mise en évidence de k événements à probabilité constante p. Y est le nombre de non événements qu'il faut observer pour pouvoir observer k événements à probabilité constante p.

Remarque : la présentation de la loi binomiale négative est parfois celle que l'on trouve ci-contre sous l'intitulé "présentation classique". Dans d'autres ouvrages, la "présentation classique" correspond à la loi de Pascal, et l'autre présentation à la loi binomiale. La loi géométrique est la loi du nombre d'essais nécessaires pour faire apparaître un événement de probabilité p.

La loi binomiale négative est aussi une distribution qui est souvent utilisée pour estimer des taux d'échantillonnage.

LOI BINOMIALE NÉGATIVE : hypothèses

sur les unités expérimentales

Y : nombre d'unités d'observation aléatoire.

sur le phénomène observé

- *événement* compté *bien défini*, *identique* pour toutes les unités expérimentales
- *comptage exhaustif* par unité expérimentale
- *un* événement compté *au plus* par unité d'observation
- *indépendance* des *unités d'observation*
- *répartition* des événements *au hasard* parmi les unités d'observation d'une unité expérimentale

Les hypothèses sont les mêmes que pour la loi de Poisson, à deux exceptions près : le nombre d'unités d'observation par unité expérimentale est aléatoire, le phénomène étudié peut être rare, mais cela n'est pas une obligation (la loi binomiale négative permet par exemple de modéliser le nombre de "pile" qu'il faut tirer pour obtenir un "face").

Attention : il faut insister sur le fait que le Y de la loi de Poisson n'a pas la même signification que le Y de la loi Binomiale Négative. Pour la loi de Poisson, il s'agit du nombre d'occurrences de l'événement pour n unités d'observation. Pour la loi Binomiale Négative, il s'agit du nombre de non événements nécessaires à l'observation de k occurrences de l'événement.

LOI BINOMIALE NÉGATIVE : variance

Distribution discrète.

$$Y' \sim \mathcal{BN}(k, p)$$

Y' : nombre d'unités d'observation qu'il faut regarder dans une unité expérimentale pour obtenir k événements, chaque événement ayant la probabilité p d'être obtenu.

$$\mathrm{E}(Y') = \frac{k}{p}$$

$$\mathrm{Var}(Y') = \frac{k(1-p)}{p^2}$$

$$\boxed{\mathrm{Var}(Y') > \mathrm{E}(Y')}$$

On peut aussi voir la relation sous la forme :

$$\mathrm{Var}(Y') = \frac{\mathrm{E}(Y')^2}{k} - \mathrm{E}(Y')$$

Distribution de probabilité pour Y' :

$$P(Y' = y) = \mathcal{C}_{y-1}^{y-k} p^k (1-p)^{y-k} \text{ pour } y \geq k$$

explication : la dernière observation est un événement. On compte donc le nombre de non-événements parmi les $y-1$ premières observations. On a k observations de l'événement, et $y-k$ observations du non-événement.

On a :

$$\mathrm{Var}(Y') = \mathrm{E}(Y')\frac{(1-p)}{p} = \mathrm{E}(Y')(\frac{1}{p} - 1)$$

et

$$\frac{1}{p} - 1 \geq 1 \Leftrightarrow p \leq 1/2$$

On peut aussi voir la relation sous la forme :

$\mathrm{Var}(Y') = \mathrm{E}(Y')^2/k - \mathrm{E}(Y')$

La loi Binomiale Négative est utilisée classiquement comme alternative empirique à la loi de Poisson quand il y a surdispersion.

LOI BINOMIALE NÉGATIVE :

autre présentation

Soit $Y = Y' - k$

Y : nombre de non-événements qu'il faut observer pour observer k événements.

Les non-événements comptés par Y ont une probabilité $(1 - p)$ constante d'être observés dans chaque essai.

$$Y \sim \mathcal{BN}(k, p)$$

$$\mathrm{E}(Y) = \frac{k}{p}(1 - p)$$

$$\mathrm{Var}(Y) = \frac{k(1 - p)}{p^2}$$

$$\boxed{\mathrm{Var}(Y) > \mathrm{E}(Y)}$$

On peut aussi voir la relation sous la forme :

$$\mathrm{Var}(Y) = \mathrm{E}(Y) + \mathrm{E}(Y)^2/k$$

Distribution de probabilité pour Y :

$$P(Y = y) = \mathcal{C}_{k+y-1}^{y} p^k (1 - p)^y \text{ pour } y \geq 0$$

On a :

$$\mathrm{Var}(Y) = \mathrm{E}(Y)/p$$

et $p < 1$.

LOI BINOMIALE NÉGATIVE

$Var(Y) = E(Y)$	loi de poisson	nb d'occurrences d'événements rares indépendants
$Var(Y) > E(Y)$	loi Binomiale Négative	nb d'occurrences d'événements non indépendants (agrégation)

La loi de Poisson est une loi discrète telle que la variance est égale à l'espérance. La loi Binomiale Négative est aussi une loi discrète, mais sa variance est supérieure à son espérance. Elle peut être employée pour des phénomènes agrégatifs.

Si l'on ne se doute pas de la nature agrégative du phénomène étudié, et que l'on utilise une loi de Poisson, on observera une surdispersion, qui doit nous mettre sur la piste de la loi Binomiale Négative.

LOI BINOMIALE NÉGATIVE

comme alternative à la loi de Poisson

- si $(Y|N = n) \sim \mathcal{P}(np)$
- si p constant
- si $N \sim$ Gamma, $\mathrm{E}(N) = \mathrm{Var}(N) = b$

alors

$$Y \sim \mathcal{BN}(b, \tfrac{1}{1+p})$$

$$\mathrm{E}(Y) = bp \qquad \mathrm{Var}(Y) = bp(p+1)$$

$$\Rightarrow \text{surdispersion} = (p+1)$$

<u>facultatif</u>

Si Y suit une distribution de Poisson conditionnellement à la taille des unités expérimentales, et si la taille des unités expérimentales suit une loi Gamma, alors Y suit une loi Binomiale Négative. La loi Binomiale Négative est donc une alternative à l'utilisation d'une loi de Poisson avec un effet aléatoire de la taille des unités expérimentales (cette solution avec effets aléatoires n'est pas simple à mettre en œuvre).

N suit une loi Gamma, on note b son espérance et sa variance. Le paramètre b est la taille moyenne des unités expérimentales.

Par rapport à une loi de Poisson, la surdispersion vaut $p+1$.

Remarque : Les unités d'observation sont toujours supposées indépendantes.

Remarque : on voit ici que la variance est proportionnelle à l'espérance.

Remarque : on voit ici que le paramètre k de la loi binomiale négative n'est pas un entier.

LOI BINOMIALE NÉGATIVE

comme alternative à la loi de Poisson

- si $(Y|P = p) \sim \mathcal{P}(np)$
- si n constant
- si $P \sim$ Gamma, $\mathrm{E}(P) = \mathrm{Var}(P) = c$

alors

$$Y \sim \mathcal{BN}(c, \tfrac{1}{1+n})$$

$$\mathrm{E}(Y) = cn \qquad \mathrm{Var}(Y) = cn(1+n)$$

$\Rightarrow$ surdispersion $= (1+n)$

<u>facultatif</u>

Considérer pour Y une distribution Binomiale Négative peut être une alternative à l'utilisation d'une distribution de Poisson avec une probabilité aléatoire suivant une loi Gamma (ce qui n'est pas simple à mettre en oeuvre).

P suit une loi Gamma, on note c son espérance et sa variance. Le paramètre c est la probabilité moyenne.

Les unités d'observation sont toujours supposées indépendantes.

La surdispersion obtenue lorsque la probabilité est aléatoire est plus grande (en général beaucoup plus grande) que la surdispersion obtenue lorsque la taille est aléatoire, car la taille n est toujours supérieure à la probabilité p.

Remarque : on voit ici que la variance est proportionnelle à l'espérance.

Remarque : on voit un nouvel exemple de loi binomiale négative où le paramètre k n'est pas entier.

LOI BINOMIALE NÉGATIVE

comme alternative à la loi de Poisson

La loi Binomiale Négative est un

modèle mathématique

qui s'ajuste très bien à des données

provenant de phénomènes

ayant un caractère "contagieux"

et à des données "agrégées" en général.

exemple :

comptage d'organismes parasites, avec
taux de croissance dépendant de l'effectif
et d'un taux d'immigration constant.

On peut utiliser la Binomiale Négative, dans le cas d'un processus de croissance d'organismes parasites, lorsque le taux de croissance dépend linéairement de l'effectif présent et d'un taux d'immigration constant ($\frac{dx}{dt} = \alpha x$, taux de croissance αx). Si le taux de croissance ne dépend que de l'immigration dont le taux est constant, on retrouve une loi de Poisson.

Autre exemple : Schukken, Casella et van den Browk[1] ont modélisé le nombre de mammites cliniques par vache en élevage bovin laitier, au moyen d'une loi Binomiale Négative (la durée de production était constante). Cela leur a permis de prendre en compte une probabilité variable d'un élevage à l'autre, même s'ils n'ont pas pu identifier les facteurs à l'origine de cette variabilité de la probabilité de mammite d'un élevage à l'autre. La modélisation leur a permis d'obtenir de très bons ajustements.

La modélisation par une loi binomiale négative leur permettait aussi de prendre en compte le caractère contagieux des mammites, bien qu'ils ne le disent pas explicitement : ils n'avaient pas les moyens de séparer les deux causes d'hétérogénéité : probabilité variable d'un élevage à l'autre, et phénomène contagieux au sein d'un élevage.

1. "Overdispersion in clinical mastitis data fron dairy herds : a negative binomial approach", Preventive Veterinary Medicine, 1991, 10: 239-245

LIENS D'APPROXIMATION ENTRE LOIS

BINOMIALE

POISSON

BINOMIALE NÉGATIVE

$$\begin{array}{ccccc} \mathcal{B}(n,p) & \longrightarrow & \mathcal{P}(np) & \longrightarrow & \mathcal{BN} \\ & \begin{cases} n \to \infty \\ p \to 0 \end{cases} & & \begin{array}{r} n \sim \text{Gamma} \\ \text{ou } p \sim \text{Gamma} \\ \text{ou agrégation} \end{array} & \end{array}$$

La loi Binomiale tend vers une loi de Poisson, lorsque $n \to \infty$ et $p \to 0$. La loi de Poisson tend vers une loi Binomiale Négative, lorsque n est aléatoire et suit une loi Gamma, ou lorsque p est aléatoire et suit une loi Gamma, où lorsque le phénomène observé est agrégatif.

Remarque : cela peut paraître surprenant de choisir une loi gamma, définie sur IR^+ pour une probabilité, bornée par 0 et 1. C'est effectivement contestable, bien que ce soit utilisé, dans des modèles de survie par exemple, où le risque suit une loi Gamma. C'est pourquoi certains auteurs partent d'une loi binomiale, avec une probabilité qui varie selon une loi Beta bornée par 0 et 1, ils obtiennent une loi Beta-Binomiale. Mais il existe très peu de logiciels qui permettent d'estimer les paramètres d'une loi Beta-Binomiale.

LOI BINOMIALE NÉGATIVE

le modèle dans le cadre GLM

Y : nombre de non-événements observés

lien log

$$\boxed{\mathrm{Var}(Y) = \mathrm{E}(Y) + \mathrm{E}(Y)^2/k}$$

Où

k est soit connu *a priori*,

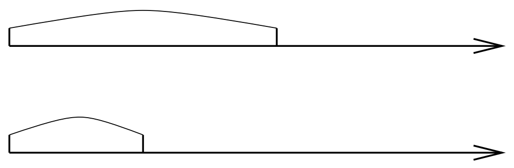
soit estimé en même temps que $\mathrm{E}(Y)$.

C'est la deuxième présentation de la Binomiale Négative qui doit être utilisée si on veut retomber sur la relation entre la variance et l'espérance ci-contre, donc celle qui fait bien intervenir la variable aléatoire Y et pas Y' ($Y = Y' - k$).

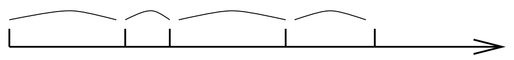
Remarque : si le paramètre k est connu, la loi binomiale négative appartient à la famille des lois exponentielles, on reste dans le cadre du modèle linéaire généralisé. Par contre, si le paramètre k est inconnu, il doit être estimé, on sort du cadre du modèle linéaire généralisé pour entrer dans le cadre de la quasi-vraisemblance.

LOI EXPONENTIELLE : présentation

La loi Exponentielle est utilisée comme modèle pour la variable aléatoire à valeur positive donnant le *temps d'attente* pour l'apparition d'un phénomène aléatoire.



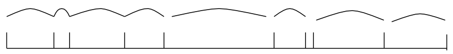
La loi Exponentielle peut servir de modèle pour une variable aléatoire à valeur positive correspondant à la *distance* séparant des événements répartis aléatoirement le long d'un itinéraire.



On a vu les lois de Poisson et Binomiale Négative, qui sont des lois discrètes. Passons à des distributions continues.

LOI EXPONENTIELLE : présentation

La loi Exponentielle s'applique si l'inverse d'une variable suivant une loi de Poisson présente une signification physique, biologique ou autre.



unité expérimentale de taille n

nombre d'événements $\sim \mathcal{P}(np)$

Y : écarts entre événements successifs.

$$Y \sim Exp(p)$$

$E(Y) = 1/p$: écart moyen entre événements

$$V(Y) = \frac{1}{p^2}$$

Si les unités expérimentales sont des intervalles de temps de taille n, si le nombre d'événements comptés suit une loi de Poisson de paramètre np, et si la variable aléatoire Y représente les écarts de temps entre les événements successifs observés, alors Y suit une distribution Exponentielle de paramètre p avec $1/p$ qui s'interprète comme l'écart de temps ou la distance moyenne entre deux événements successifs.

LOI GAMMA : présentation

La loi Gamma

est une généralisation de la loi Exponentielle

elle est utilisée comme modèle

pour la variable aléatoire à valeur positive

donnant le temps séparant

l'apparition de deux événements

séparés par un nombre fixé d'événements.

Quand les deux événements sont successifs

on retrouve la loi Exponentielle.

La loi Gamma est appliquée comme modèle de probabilité pour prévoir la durée de vie des appareils qui subissent une usure, tels les véhicules automobiles ou les appareils mécaniques.

Remarque : la loi Gamma est plutôt utilisée dans la pratique comme modèle général à caractère théorique, et plus rarement comme modèle pratique d'explication de phénomènes biologiques. Par exemple la loi du χ^2 ("khi deux" ou plutôt "khi carré") est un cas particulier de la loi Gamma. Elle peut aussi être vue comme la loi de la somme des carrés de variables aléatoires indépendantes qui suivent une loi Gaussienne centrée et réduite.

LOIS GAMMA ET EXPONENTIELLE : hypothèses

Nombre d'événements $\sim \mathcal{P}(np)$

$$\Longleftrightarrow$$

écart entre événements successifs

$$\begin{array}{rcl} T_l - T_{l-1} & \sim & \text{Exponentielle}(p) \\ T_1 & \sim & \text{Exponentielle}(p) \end{array}$$

$$\Longleftrightarrow$$

$$\begin{array}{rcl} T_l - T_{l-k} & \sim & \text{Gamma}(p,k) \\ T_k & \sim & \text{Gamma}(p,k) \end{array}$$

donc

mêmes hypothèses que pour la loi de Poisson

$T_l - T_{l-1}$ est l'écart entre la l ième et la l-1 ième apparition du phénomène. T_1 est le temps d'attente pour la 1 ière apparition du phénomène.

$T_l - T_{l-k}$ est l'écart entre la l ième et la l-k ième apparition du phénomène. T_k est le temps d'attente pour la kième apparition du phénomène.

La loi Gamma peut être présentée à partir de la loi de Poisson. Les hypothèse sont les mêmes que pour la loi de Poisson.

LOI GAMMA : variance

Distribution continue

$$Y \sim \text{Gamma}(p, k)$$

$$\text{E}(Y) = \frac{k}{p} \qquad \text{Var}(Y) = \frac{k}{p^2}$$

$$\boxed{\text{Var}(Y) = [\text{E}(Y)]^2/k}$$

$$\text{Gamma}(p, 1) = \text{Exponentielle}(p)$$

densité : $f(y) = \frac{p(py)^{k-1}}{(k-1)!} \exp(-py)$ pour $y \geq 0$ et $k \geq 1$.

Interprétation en terme d'occurrence d'événements si Y est une durée.

Soient T_k l'instant d'arrivée du k ième événement et D_k l'intervalle de temps séparant le k ième événement du $(k-1)$ ième, les intervalles D suivent une distribution Exponentielle de paramètre p et T_k suit une distribution Gamma de paramètres p et k. De même le temps écoulé entre le r ième et le $(r+k)$ ième suit aussi une distribution Gamma de paramètre p et k. On a le même type d'interprétation en terme de distance.

Cas particulier : Avec $p = 1$, on obtient la formulation de la loi Gamma de paramètre k, vue lors de l'approximation de la loi de Poisson par une loi Binomiale Négative, soit :

$f(y) = \frac{y^{k-1}}{(k-1)!} \exp(-y)$ pour $y \geq 0$ et $k \geq 1$

$\text{E}(Y) = k$ et $\text{Var}(Y) = k$

On généralise en introduisant la fonction Gamma Γ dans les formules car $\Gamma(k) = (k-1)!$ quand k est entier, puis

$\Gamma(k) = \int_0^\infty y^{k-1} \exp(-y) dy$ pour k quelconque tel que $k > 0$. Dans ce cas, le paramètre k devient la taille moyenne de la variable aléatoire N.

Généralisation de la loi Gamma de paramètres p et k (valeurs strictement positives quelconques) :

$f(y) = \frac{p(py)^{k-1}}{\Gamma(k)} \exp(-py)$ pour $y \geq 0$

$\text{E}(Y) = \frac{k}{p} \qquad \text{Var}(Y) = \frac{k}{p^2}$

Remarque : quand k est entier, la loi Gamma est aussi appelée loi d'Erlang.

Remarque : quand $p = \frac{1}{2}$ et $k = \frac{1}{2}\nu$ avec $\nu > 0$, pour tout $y \geq 0$, la loi Gamma est aussi la loi du χ^2 ("khi carré") et le paramètre ν s'interprète en terme de degrés de liberté.

LOIS GAMMA ET EXPONENTIELLE

le modèle dans le cadre GLM

loi continue

Y : temps ou distance,

ou autre.

lien canonique : fonction “inverse”,

$$\boxed{\mathrm{Var}(Y) = \mathrm{E}(Y)^2/k}$$

Les lois gamma et exponentielle sont des lois continues, souvent utilisées lorsque la variable aléatoire étudiée mesure un temps ou une distance, mais aussi dans d'autres cas. Elles sont utiles pour modéliser des distributions continues, disymétriques, lorsque la variance est proportionnelle au carré de l'espérance. Elles permettent, par exemple, de modéliser des hauteurs de précipitation, ou encore la durée du développement embryonnaire de la drosophile.

Dans le cas particulier où $k = 1$, la variance est égale au carré de l'espérance.

AUTRES LOIS : récapitulatif

lois discrètes :

Binomiale	$\mathrm{Var}(Y) < \mathrm{E}(Y)$
Poisson	$\mathrm{Var}(Y) = \mathrm{E}(Y)$
Binomiale Négative	$\mathrm{Var}(Y) > \mathrm{E}(Y)$

lois continues :

Gaussienne	$\mathrm{Var}(Y) = \text{Constante}$
Exponentielle	$\mathrm{Var}(Y) = \mathrm{E}(Y)^2$
Gamma	$\mathrm{Var}(Y) = \mathrm{E}(Y)^2/k$

quasi-vraisemblance :

modéliser la relation entre $\mathrm{Var}(Y)$ et $\mathrm{E}(Y)$

sans faire le choix d'une loi,

méthode utilisée en cas de surdispersion.

Les lois binomiale, poisson, gaussienne, exponentielle et gamma font partie de la famille des lois exponentielle. Elles relèvent donc du modèle linéaire généralisé. La loi binomiale négative fait partie de la famille des lois exponentielles si le paramètre k est connu. Dans ce cas, elle relève du modèle linéaire généralisé.

Par contre, lorsqu'on ne choisit pas une loi de la famille exponentielle, ou lorsqu'on ne choisit pas de loi du tout et qu'on se contente de modéliser la liaison entre variance et espérance, on n'utilise plus la méthodologie du modèle linéaire généralisé, mais une extension. L'estimateur n'est plus un estimateur du maximum de vraisemblance. On maximise une fonction qui n'est plus la vraisemblance, on parle de quasi-vraisemblance.

C'est le cas lorsqu'on a surdispersion. La loi n'est plus une loi de la famille exponentielle, la méthode relève de la quasi-vraisemblance, on ne fait que modéliser la liaison entre variance et espérance.

résultats généraux

$$E(Y) = \mu$$

$$\boxed{g(\mu) = X\theta = \eta \qquad \text{loi de } Y?}$$

loi	lien canonique	fct. de var. V
Normale	μ	1
Binomiale	$\log(\frac{\mu}{1-\mu})$	$\mu(1-\mu)$
Poisson	$\log(\mu)$	μ
Gamma	$\frac{1}{\mu}$	μ^2
Gaussienne inverse	$\frac{1}{\mu^2}$	μ^3
quelconque	$g(\mu)$	$V(\mu)$

	logit	probit	cloglog	id.	log	sqrt	$\frac{1}{\mu}$	$\frac{1}{\mu^2}$
Normale				D				
Binomiale	D	•	•					
Poisson				•	D	•		
Gamma				•	•		D	
Gauss. inv.								D
quelconque	•	•	•	•	•	•	•	•

La loi de Y choisie, les résultats généraux permettent de calculer espérance et variance.

Le choix de la fonction de lien découle directement de la formalisation du problème. Dans les courbes dose-réponse, il découle de la distribution supposée des seuils de tolérance. De manière générale, on recherche le lien correspondant à une transformation qui linéarise la forme de l'effet, et rend les effets additifs.

Insister sur le fait qu'en GLM, on dispose d'**un outil** pour obtenir des **effets additifs et tels que décrits** dans la partie systématique du modèle, c'est la **fonction de lien**; et d'**un autre outil**, pour **modéliser l'hétérogénéité des variances**, c'est le **choix d'une loi** pour Y.

Pour chaque loi, il existe un lien par défaut (noté D dans le tableau), dit lien canonique, qui possède de bonnes propriétés statistiques, mais il est possible de choisir un autre lien qui conviendrait mieux (le lien "id." est le lien identité).

Rq formateur : les liens présentés dans le tableau sont les liens les plus fréquemment utilisés, mais on peut en inventer d'autres.

Rq : quand la loi est quelconque, quand elle n'appartient pas à la famille des lois exponentielles, ou si ϕ est inconnu, on utilise une extension du GLM mettant en oeuvre la notion de **quasi-vraisemblance** qui permet **la modélisation de la variance, sans spécifier la loi**. Les lois normale, binomiale, de poisson, gamma, gaussienne inverse appartiennent à la famille des lois exponentielles.

Rq : la **loi multinomiale** (extension de la loi binomiale à plus de deux modalités) rentre aussi **dans le cadre du modèle linéaire généralisé**. Elle sera abordée au chapitre VII.

Généralisation : la famille exponentielle

- densité $f(y;\theta,\phi) \in$ famille exponentielle si :

$$\log(f(y;\theta,\phi)) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)$$

- θ : **paramètre canonique**
- ϕ : paramètre de dispersion (connu)
- $E(Y) = \mu = b'(\theta)$
- $Var(Y) = b''(\theta)a(\phi)$
- $b''(\theta)$ **fonction de variance**

Ce transparent, ainsi que les trois transparents qui suivent sont optionnels. Dans toute la première partie du cours, on s'est attaché à introduire le modèle linéaire généralisé dans le cas particulier de la loi binomiale. On généralise maintenant à la famille des lois exponentielles, à laquelle appartient la loi binomiale. On reste ainsi dans le cadre du modèle linéaire généralisé.

Attention au changement de notation : θ est ici une fonction de μ. $\theta = g(\mu) = \eta$. Ce θ n'a rien à voir avec le θ utilisé jusqu'à présent, qui était le vecteur des paramètres du modèle. On a conservé la notation θ dans la présentation de la famille exponentielle, parce que c'est la notation traditionnellement utilisée.

Le modèle linéaire généralisé explique θ par une combinaison linéaire des prédicteurs. Pour les lois classiques de la famille exponentielle (binomiale, normale, Poisson, Gamma,...) θ est une fonction de l'espérance μ de ces lois. Cette fonction est appelée **lien canonique**.

ϕ est souvent noté σ^2. ϕ est supposé connu.

Lorsque ϕ est inconnu, c'est à dire lorsqu'on estime un paramètre de dispersion, l'estimateur n'est plus un estimateur du maximum de vraisemblance, et la méthode employée ne relève plus du modèle linéaire généralisé, mais d'une extension du modèle linéaire généralisé (la quasi-vraisemblance).

θ étant une fonction de μ, on note $V(\mu)$ la fonction de variance.

application : **loi binomiale**
(loi d'une proportion X)

- $f(x;p) = \mathcal{C}_n^{nx}\, p^{nx}(1-p)^{n-nx}$
  $\log f(x;p) = n(x \log \dfrac{p}{1-p} + \log(1-p)) + \log \mathcal{C}_n^{nx}$
- $\theta = \log(\frac{p}{1-p})$ lien canonique entre p et θ
- $b(\theta) = -\log(1-p) = \frac{1}{1+\exp\theta}$
- $\phi = 1/n,\ a(\phi) = \phi$
- $c(x,\phi) = \log(\mathcal{C}_n^x)$
- $E(X) = \mu = \dfrac{\exp\theta}{1+\exp\theta} = p$
- $Var(X) = b''(\theta)a(\phi) = \mu(1-\mu)/n$
  $V(\mu) = \mu(1-\mu)$ fonction de variance

Rappel : quelques formules de dérivée.

$$(\frac{u}{v})' = \frac{u'v - uv'}{v^2}$$

$$(\log(x))' = \frac{1}{x}$$

$$(\frac{1}{x})' = -\frac{1}{x^2}$$

$$(\exp x)' = \exp(x)$$

$$[g(f(u))]' = g'(f(u))f'(u)$$

application : **loi normale**

- $f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-(x-\mu)^2/2\sigma^2)$

- $\log f(x;\mu,\sigma^2) = \frac{x\mu-\mu^2/2}{\sigma^2} - 1/2(x^2/\sigma^2 + \log(2\pi\sigma^2))$

- $\theta = \mu$ **lien canonique** (identité)

- $b(\theta) = \theta^2/2$

- $\phi = \sigma^2$, $a(\phi) = \phi = \sigma^2$

- $c(x,\phi) = -1/2(x^2/\sigma^2 + \log(2\pi\sigma^2))$

- $E(X) = \mu$

- $Var(X) = b''(\theta)a(\phi) = 1 \times \sigma^2$
  $V(\mu) = 1$ **fonction de variance**

Si surdispersion

$$V(Y) = \sigma^2 V(\mu)$$

σ^2 inconnu

on n'est plus dans le cadre GLM,

maximisation d'une fonction $\rightarrow \widehat{\theta}$

quasi-vraisemblance

En cas de surdispersion, la loi n'est plus une loi de la famille exponentielle, la méthode relève de la quasi-vraisemblance. On ne fait que modéliser la liaison entre variance et espérance. L'estimateur de quasi-vraisemblance n'est pas défini comme réalisant l'optimum d'un critère, mais comme solution d'équations d'estimation.

Si surdispersion

$$V(Y) = \sigma^2 V(\mu)$$

σ^2 paramètre de surdispersion

$$\boxed{D \sim_{as} \sigma^2 \chi^2(m-k)}$$

$$\boxed{D_2 - D_1 \sim_{as} \sigma^2 \chi^2(dl_2 - dl_1)}$$

1. $\sigma^2 = 1$: tests du χ^2
2. σ^2 inconnu : tests de Fisher

$$\frac{(D_2 - D_1)/(dl_2 - dl_1)}{D_0/dl_0} \sim F(dl_2 - dl_1, dl_0)$$

Lorsque σ^2 est inconnu, c'est à dire lorsqu'on estime un paramètre de dispersion, l'estimateur n'est plus un estimateur du maximum de vraisemblance, et la méthode employée ne relève plus du modèle linéaire généralisé, mais d'une extension du modèle linéaire généralisé. La déviance est définie comme au chapitre 3, avec un σ^2 inconnu.

La loi asymptotique de la déviance n'est plus un χ^2, mais un $\sigma^2\chi^2$. $\boldsymbol{\sigma^2}$ est un **paramètre de surdispersion**, parfois appelé **paramètre d'échelle** (scale parameter). On en déduit que $D_2 - D_1 \sim_{as} \sigma^2\chi^2(dl_2 - dl_1)$.

- Dans le cas binomial, σ^2 vaut 1 ($var(\frac{Y_i}{n_i}) = \sigma^2 \frac{p_i(1-p_i)}{n_i}$, avec $\sigma^2 = 1$).

  Dans le cas de la loi de poisson, σ^2 est aussi fixé égal à 1 ($Var(Y) = \mu$). Si $\sigma^2 = 1$, la déviance est une mesure d'adéquation du modèle, elle suit, sous l'hypothèse d'un modèle adéquat, une loi $\chi^2(m-k)$. Résultat peu valide dans le cas de la loi binomiale, avec des $n_i = 1$ ou trop petits.

- Lorsqu'il y a **sur ou sousdispersion**, un paramètre σ^2 est estimé, différent de 1. La déviance dépend de σ^2, elle n'est plus une mesure d'adéquation du modèle. La différence de deux déviances dépend de σ^2, elle ne peut être une statistique de test pour comparer deux modèles. Par analogie avec le modèle linéaire gaussien, la statistique $\frac{(D_2-D_1)/(dl_2-dl_1)}{D_0/dl_0}$, où D_0 est la déviance du modèle le plus complet, ne dépend pas de σ^2, elle est utilisée pour réaliser des tests de sous-modèle. **Les tests de sous-modèle sont réalisés au moyen de tests de Fisher**, en prenant comme modèle "de référence" le plus grand modèle (on appelle modèle de référence, le modèle qui sert à estimer la variabilité résiduelle qui intervient au dénominateur de la statistique de Fisher).

  Remarquons que le modèle M_1 associé à D_1 doit être différent du modèle M_0 associé à D_0. Sinon, cela reviendrait à faire un test d'adéquation, et ce test n'est pas valide.

Si surdispersion

Maximisation d'un critère $\Rightarrow \widehat{\theta}$

$$\boxed{\widehat{\theta} \sim_{as} \mathcal{N}(\theta, \sigma^2 (X'WX)^{-1})}$$

σ^2 remplacé par $\widehat{\sigma}^2$

$\rightarrow$ Test de nullité d'un paramètre

$$\frac{\widehat{\theta}_j}{\widehat{se(\widehat{\theta}_j)}} \sim_{as} t(dl)$$

$\widehat{\theta} \sim_{as} \mathcal{N}(\theta, \sigma^2(X'WX)^{-1})$ (W matrice de pondération). σ^2 est inconnu, on le remplace par $\widehat{\sigma}^2$, ce qui permet de tester la nullité d'un paramètre. Pratiquement, les écarts-type des paramètres obtenus en faisant l'hypothèse d'absence de surdispersion sont multipliés par $\widehat{\sigma}$, et le test est réalisé comme on réalise un test de Wald (avec Splus, c'est ce que fait la fonction summary, avec l'option dispersion=0).

Cela explique qu'en général, les logiciels nomment "t-value" les $\frac{\widehat{\theta}_j}{\widehat{se(\widehat{\theta}_j)}}$, et non pas z-value. Il faut considérer que les quantités affichées dans la colonne intitulée t-value sont en fait des z-value dans tous les cas où il n'y a pas surdispersion.

Dans tous les cas où un paramètre de variance est estimé,

$$\frac{\widehat{\theta}_j}{\widehat{se(\widehat{\theta}_j)}} \sim_{as} t(dl)$$

où dl est le degré de liberté résiduel.

C'est le cas lorsqu'on estime un paramètre σ^2 de surdispersion.

C'est parce qu'on a ajouté un chapeau sur le "se" que la statistique suit une loi de student (dont le degré de liberté est le nombre de degrés de liberté résiduels), et non pas une loi normale.

Analogie entre LM, GLM et quasi-vraisemblance :

<u>LM :</u>	<u>GLM :</u>	<u>Surdispersion :</u>
$V(\varepsilon_i) = \sigma^2$	$V(Y) = V(\mu)$	$V(Y) = \sigma^2 V(\mu)$
$\theta \sim \mathcal{N}(\theta, \sigma^2(X'X)^{-1})$	$\widehat{\theta} \sim_{as} \mathcal{N}(\theta, (X'WX)^{-1})$	$\widehat{\theta} \sim_{as} \mathcal{N}(\theta, \sigma^2(X'WX)^{-1})$
$\frac{\widehat{\theta}-\theta}{\sigma} \sim \mathcal{N}(0,1)$	$\frac{\widehat{\theta}-\theta}{se(\widehat{\theta})} \sim_{as} \mathcal{N}(0,1)$	$\frac{\widehat{\theta}-\theta}{\widehat{se(\widehat{\theta})}} \sim_{as} t$
$\frac{\widehat{\theta}-\theta}{\widehat{\sigma}} \sim t$		