Recensement de l'invisible

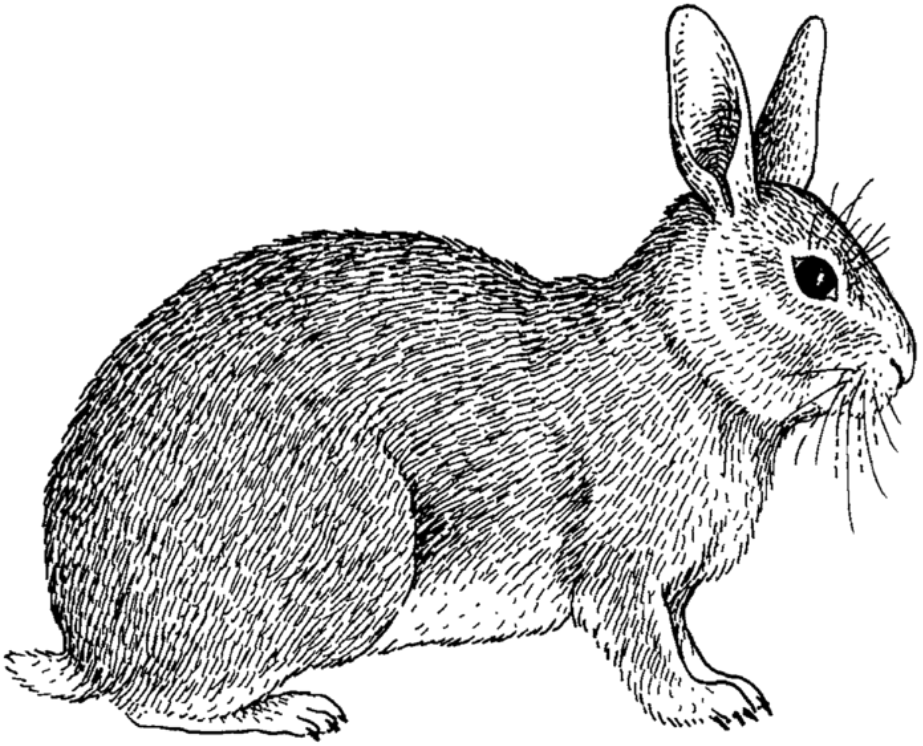
Hervé Perdry - 28 janvier 2013

Séminaire Quali
Quanti

Sommaire

Comment compter plusieurs choses que je n'ai pas observées :

- des lapins (capture-recapture)
- des porteurs d'une mutation d'un gène suppresseur de tumeur (loi de Poisson tronquée)
- des espèces de papillons ou... le vocabulaire de Proust ?



(Lecointre et Le Guyader)

Modèle de capture - recapture

Problème : combien y a-t-il de lapins dans le petit bois derrière chez moi ?

La solution proposée est la suivante (méthode de Lincoln-Petersen) :

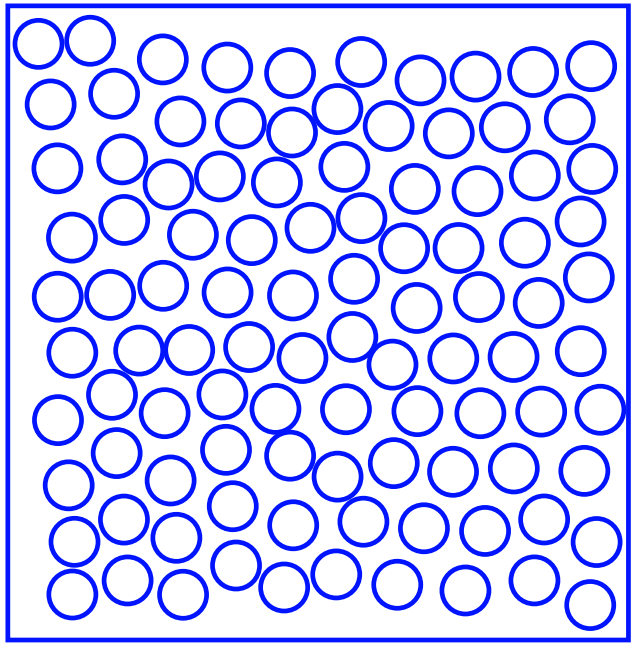
- on note N le nombre de lapins ;
- (capture) lundi, on piège n lapins, on les marque, on les relâche ; une proportion $p = n/N$ de lapins sont désormais marqués ;
- (recapture) vendredi, on repiège des lapins ; une partie d'entre eux est marquée ;
- on estime p par $\widehat{p}$, la proportion de lapins marqués qu'on observe lors de la recapture ;
- on estime N par $n/\widehat{p}$.

☞ Méthode très utilisée en écologie, mais aussi pour calculer l'effectif de populations humaines difficiles à observer, en particulier usagers de drogues, etc.

Modèle de capture - recapture

Exemple

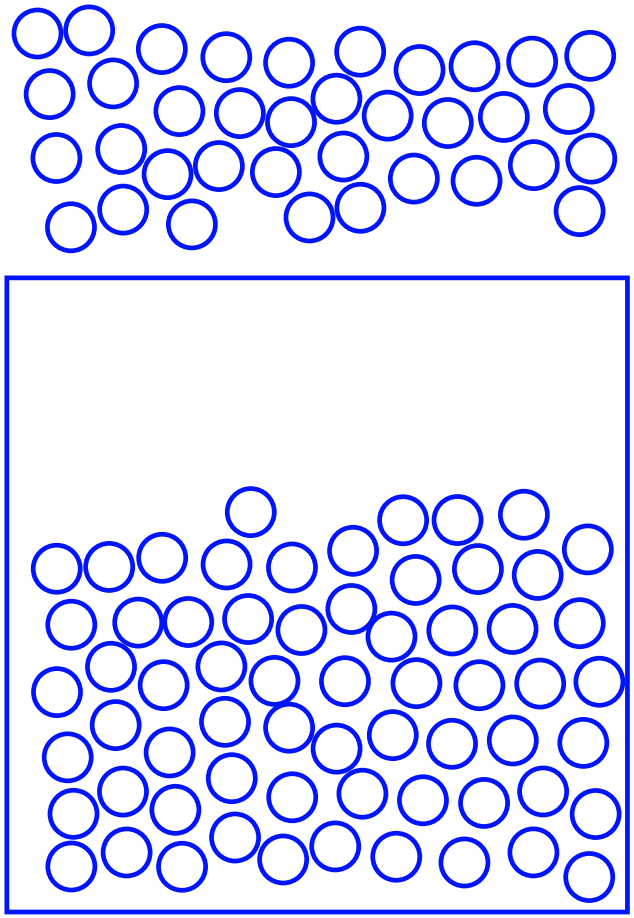
On a une urne qui contient un nombre inconnu de boules.



Modèle de capture - recapture

Exemple

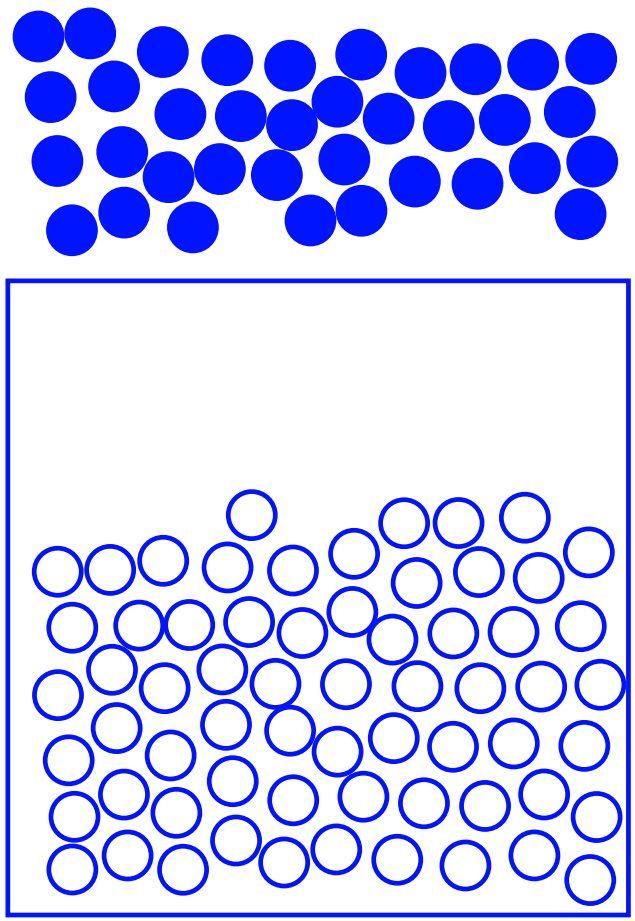
On en extrait $n = 36$ boules...



Modèle de capture - recapture

Exemple

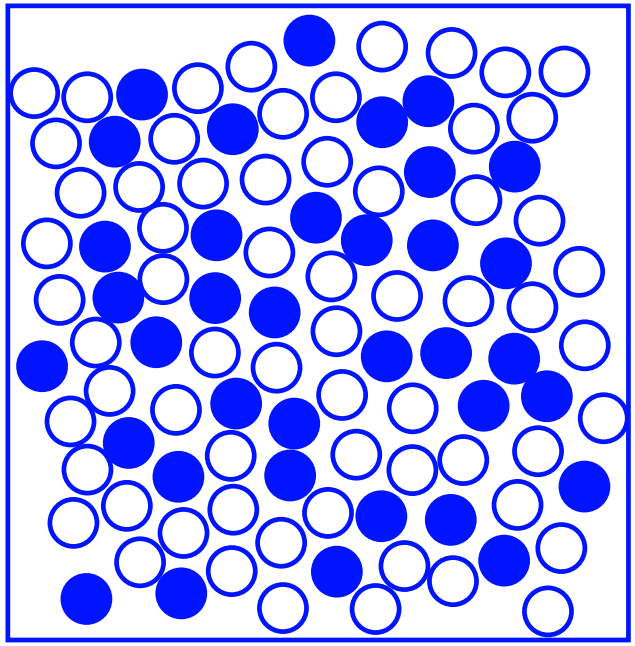
...qu'on marque.



Modèle de capture - recapture

Exemple

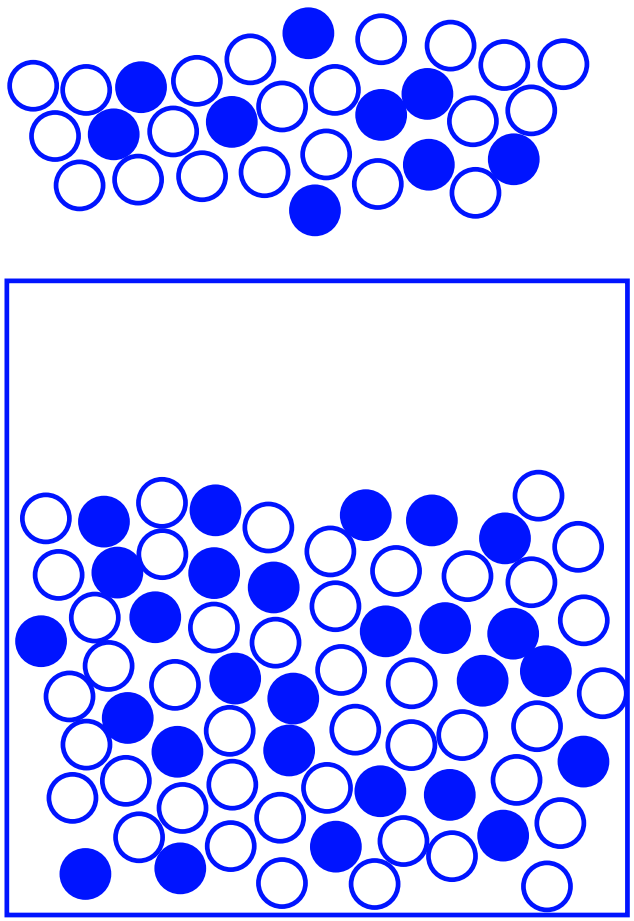
On les remet dans l'urne et... on mélange !



Modèle de capture - recapture

Exemple

On tire 30 boules dans l'urne et on compte les boules marquées : il y en a 9.



Modèle de capture - recapture

Exemple

Bilan :

- on a marqué $n = 36$ individus
- au second tirage, on estime qu'une proportion $\widehat{p} = \frac{9}{30} = 0,3$ d'individus sont marqués
- on a $p = 36/N$ d'où on estime $\widehat{N} = \frac{36}{0,3} = 120$

On peut aussi obtenir un intervalle de confiance à 95 %...

$$\widehat{p} \in \left[0,3 \pm 1,96 \sqrt{\frac{0.3 \times 0.7}{30}} \right]$$

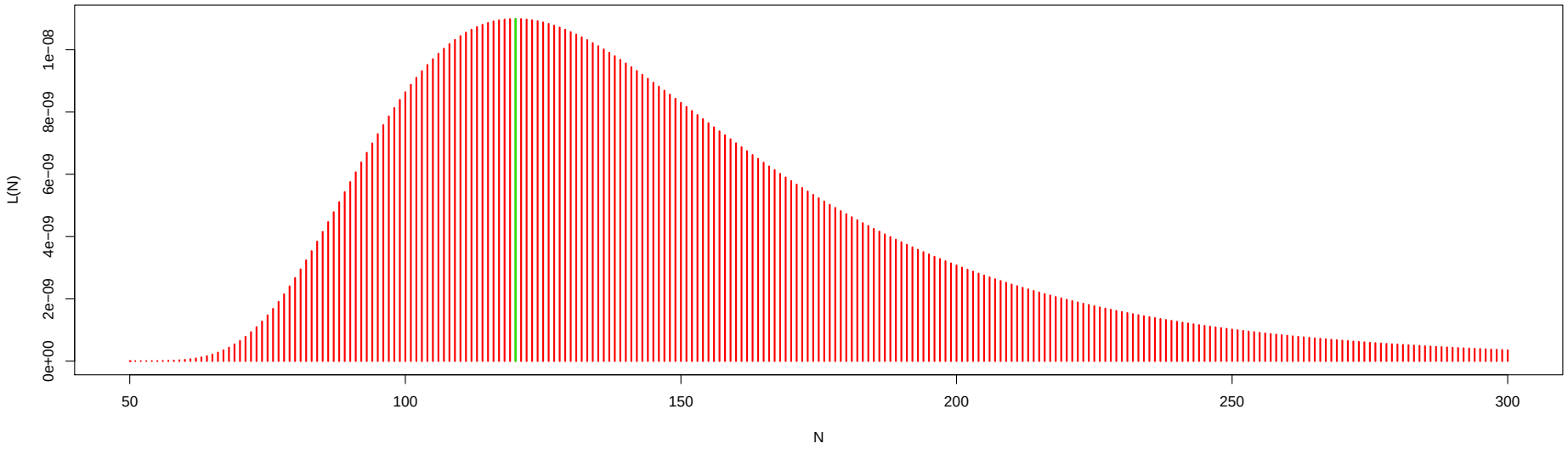
c'est à dire p entre 0,136 et 0,464, d'où N entre 78 et 264...

Modèle de capture - recapture

Avec une vraisemblance...

On peut aussi écrire la vraisemblance des observations faites à la seconde campagne de capture, comme une fonction de N :

$$L(N) = \left(\frac{36}{N}\right)^9 \times \left(1 - \frac{36}{N}\right)^{21}$$

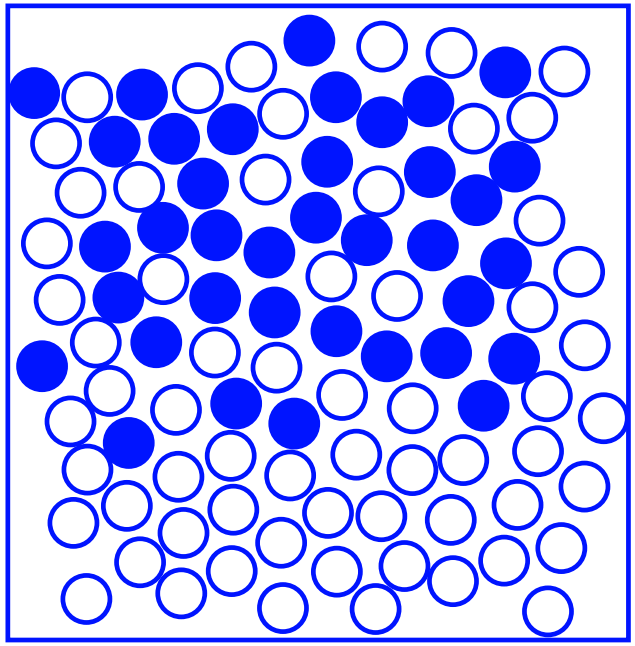


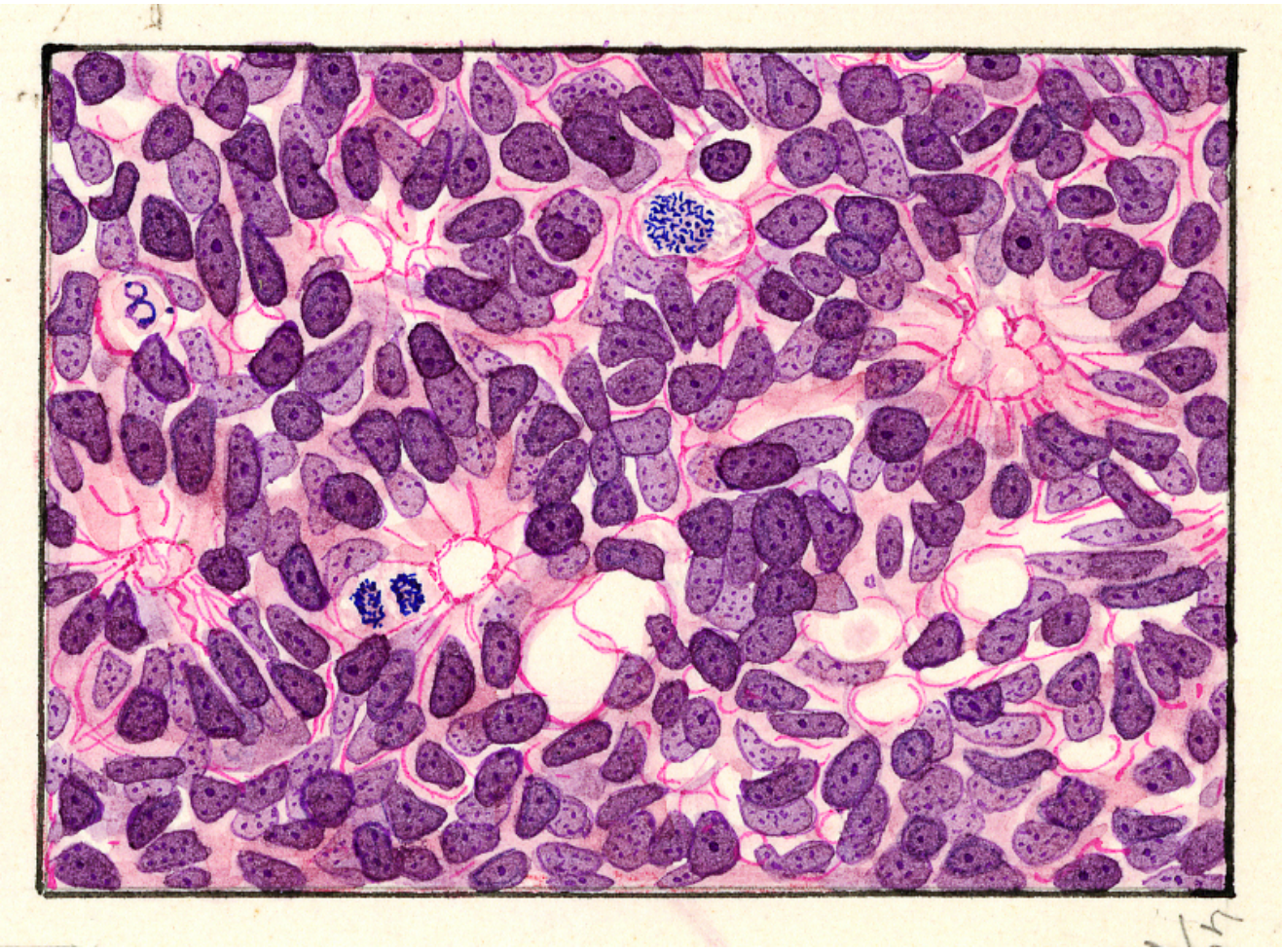
Modèle de capture - recapture

Commentaires

Problème... et si on n'a pas bien mélangé l'urne ? Les individus marqués sont plus susceptibles d'être recapturés que les autres...
(beaucoup d'autres hypothèses / d'autres problèmes possibles).

☞ Il existe des quantités de modèles plus complexes (campagnes multiples, modélisation de la structure géographique, de la mortalité, etc)





Dessin de Jacques Mawas

Rétinoblastome

Contexte

- Cancer de la rétine chez l'enfant (avant 5 ans)
- Maladie dominante due à une mutation dans un gène récepteur de tumeur (RB1)
- Si une seconde mutation apparaît pendant le développement de l'oeil, une tumeur apparaît
- Cet événement peut se produire plusieurs fois : tumeurs multiples

Rétinoblastome

Contexte

Nombre de tumeurs observées chez des patients (qui portent tous la mutation) :

Nb tumeurs k	Nb patients S_k
1	35
2	17
3	9
4	4
5	1

117 tumeurs, 66 patients (Knudson 1971)

- Quel modèle pour ces données ?
- Combien d'individus (dans la zone d'origine de ces patients) portent la mutation mais n'ont pas développé de tumeur ?

Rétinoblastome

Modèle

Si les tumeurs surviennent indépendamment les unes des autres pendant une période de temps limitée, on a un processus de Poisson ; le nombre de tumeurs suit une loi de Poisson :

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

On a $E(X) = \mathrm{Var}(X) = \lambda$.

Rétinoblastome

Estimer s (λ connu)

Si s est le nombre d'individus qui portent la mutation, on attend $s \times \mathbb{P}(X = k)$ patients avec k tumeurs.

L'espérance du nombre d'individus avec au moins une tumeur est

$$E(S_1 + S_2 + \cdots) = s \times \mathbb{P}(X > 0) = s \times (1 - e^{-\lambda})$$

On peut estimer s par

$$\boxed{\widehat{s} = \frac{\sum_{k \geq 1} S_k}{1 - e^{-\lambda}}}$$

Rétinoblastome

Estimer λ

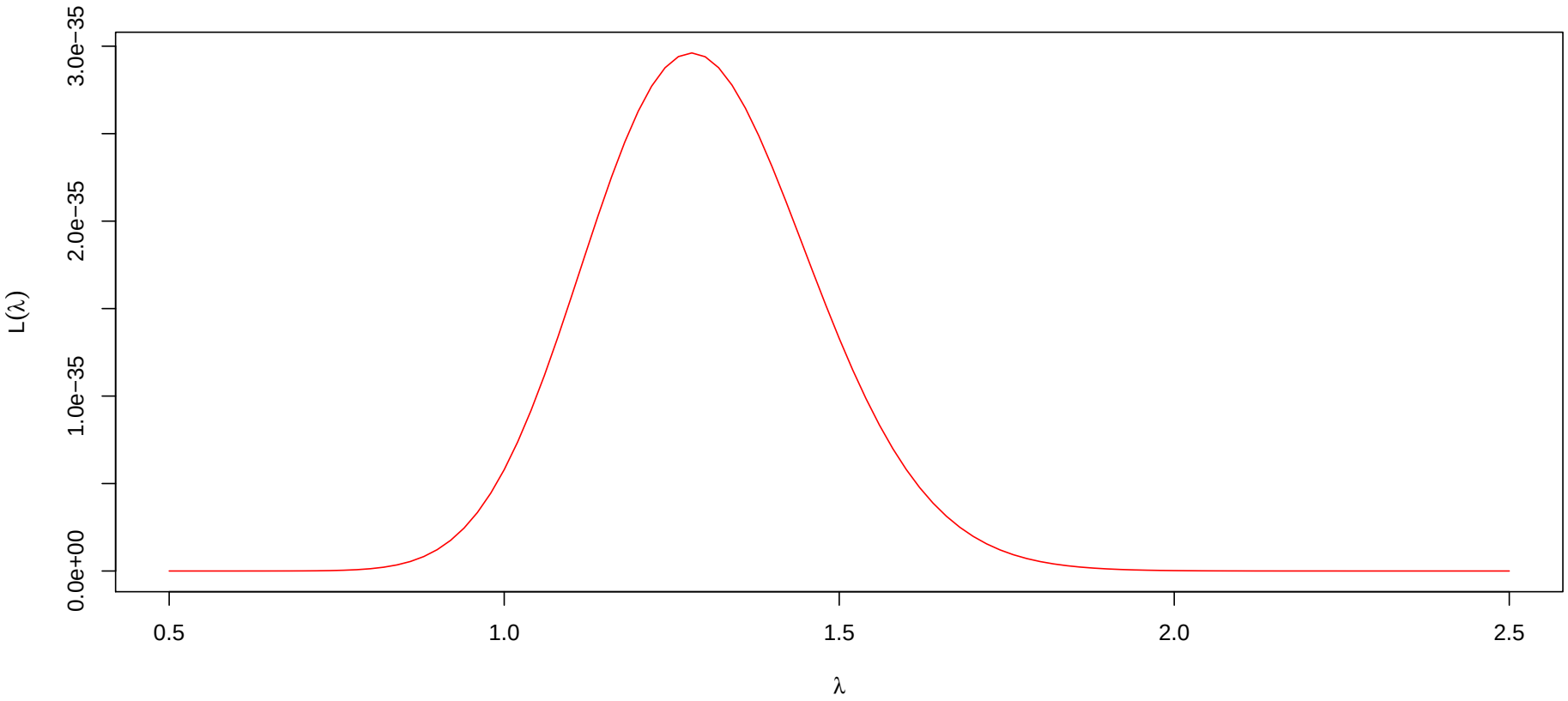
Comme on n'observe pas S_0, on s'intéresse à la probabilité d'observer une valeur $X = k$ conditionnellement à $X > 0$:

$$\begin{aligned}\mathbb{P}(X = k | X > 0) &= \frac{\mathbb{P}(X = k)}{1 - \mathbb{P}(X = 0)} \\ &= \frac{\lambda^k}{k!} \frac{e^{-\lambda}}{1 - e^{-\lambda}}\end{aligned}$$

Rétinoblastome

Estimer λ : la vraisemblance

$$L(\lambda) = \prod_k \mathbb{P}(X = k | X > 0)^{S_k}$$



Rétinoblastome

Estimer λ : la vraisemblance

Le maximum de vraisemblance est pour $\lambda = 1,28$.
On en déduit que la probabilité de n'avoir aucune tumeur chez un porteur est

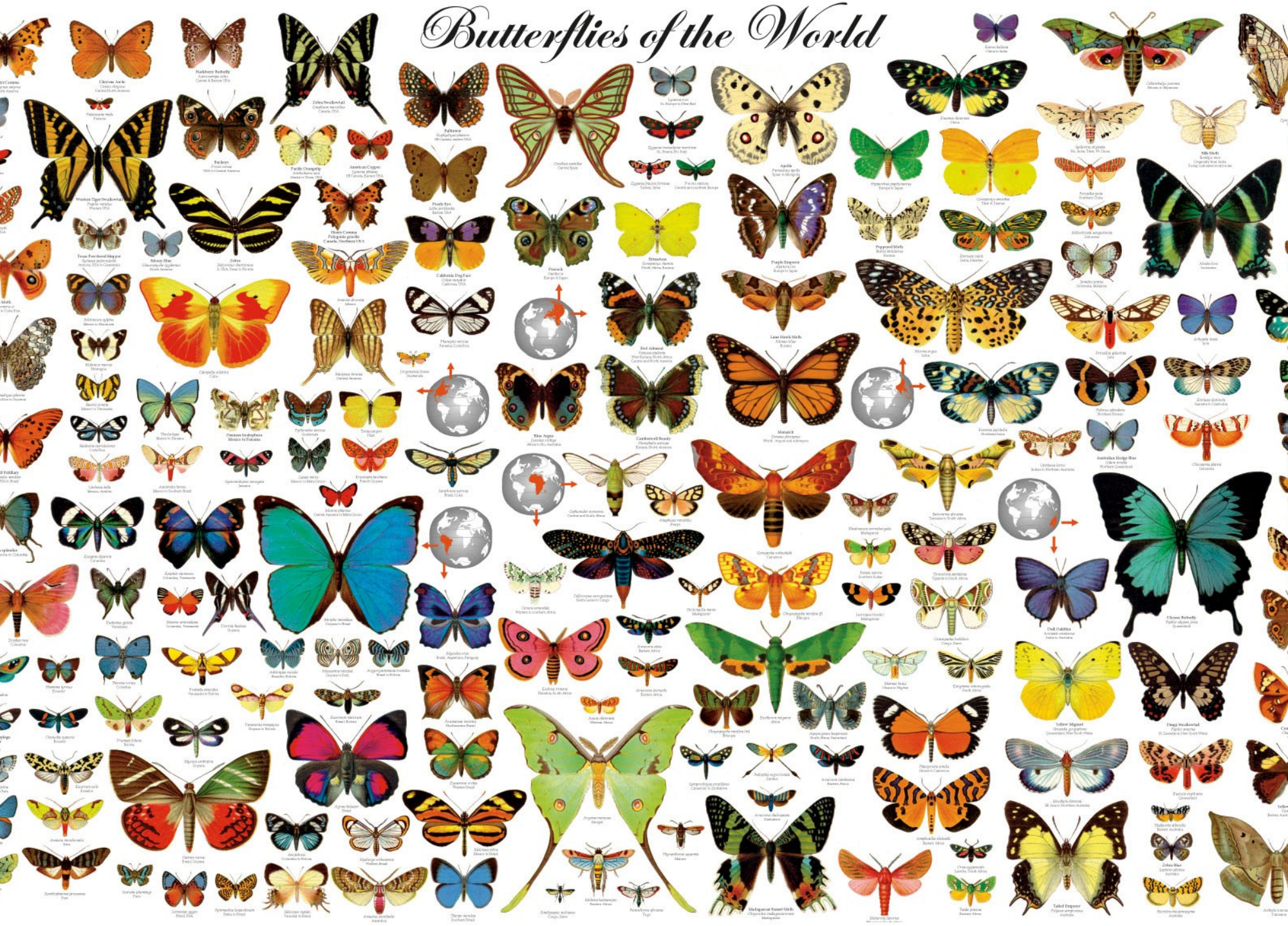
$$1 - e^{1,28} = 0,72$$

et on estime le nombre total de porteur dans la zone de recrutement par

$$\frac{66}{0,72} \simeq 91$$

Balayé sous le tapis : il s'agit en fait du nombre de tumeurs dans un des deux yeux chez des patients avec atteinte bilatérale ; il faudrait prendre $\lambda = 2 \times 1,28$ pour faire le calcul ci-dessus, le nombre total de tumeurs (dans les deux yeux) suivant une loi $\mathscr{P}(2\lambda)$. Et tant qu'on y est, il faudrait tenir compte du fait que l'œil où le médecin a pu dénombrer les tumeurs est probablement le moins atteint des deux ! (et de l'âge du patient...)

Butterflies of the World



Compter les papillons

Contexte

On capture des papillons, et on les classe par espèces. Certaines espèces (beaucoup !) sont observées $k = 1$ seule fois, d'autres $k = 2$ fois, etc.

k	S_k	k	S_k	k	S_k	k	S_k	k	S_k	k	S_k
1	118	5	29	9	20	13	6	17	9	21	11
2	74	6	22	10	15	14	12	18	6	22	5
3	44	7	20	11	12	15	6	19	10	23	3
4	29	8	19	12	14	16	9	20	10	24	3

Répartition en 501 espèces de 3306 papillons capturés en Malaisie

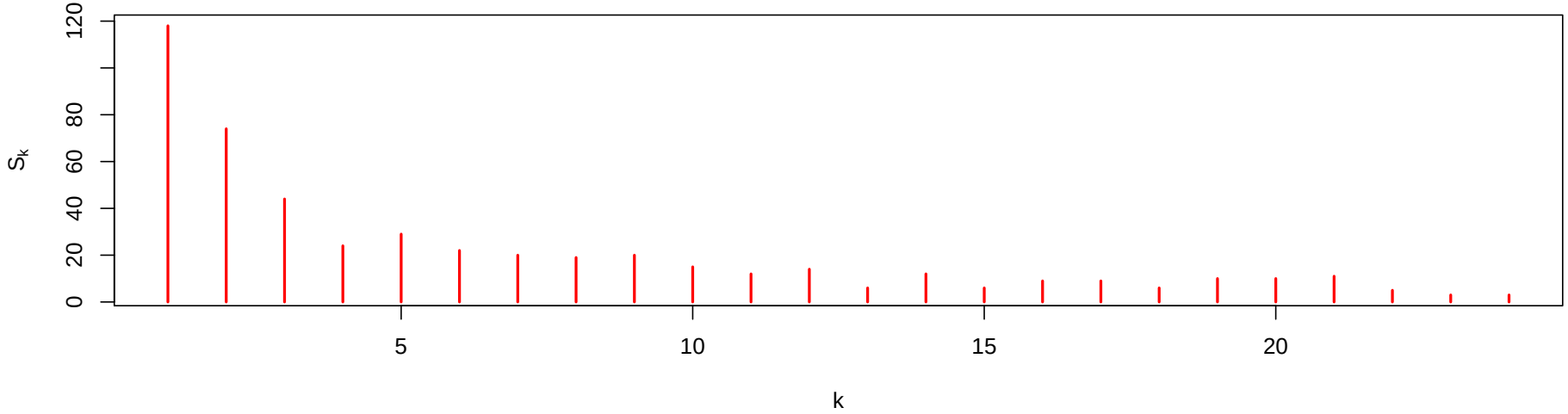
Questions :

- Donner un modèle qui « explique » ce type de répartition
- Combien d'espèces nouvelles seraient découvertes par une nouvelle campagne de capture ?
- Estimer le nombre d'espèces à découvrir (? !)

Compter les papillons

Contexte

On capture des papillons, et on les classe par espèces. Certaines espèces (beaucoup !) sont observées $n = 1$ seule fois, d'autres $n = 2$ fois, etc.



Questions :

- Donner un modèle qui « explique » ce type de répartition
- Combien d'espèces nouvelles seraient découvertes par une nouvelle campagne de capture ?
- Estimer le nombre d'espèces à découvrir (? !)

Compter les papillons

Modèle

- Le nombre X_i d'individus capturés dans une espèce i donnée suit une loi de Poisson de paramètre λ_i (inconnu) propre à l'espèce :

$$X_i \sim \mathscr{P}(\lambda_i)$$

c'est-à-dire que

$$\mathbb{P}_{\lambda_i}(X_i = k) = \frac{\lambda_i^k}{k!} e^{-\lambda_i}.$$

- Les λ_i sont pris « au hasard » dans une loi $\Gamma(r, \theta)$.

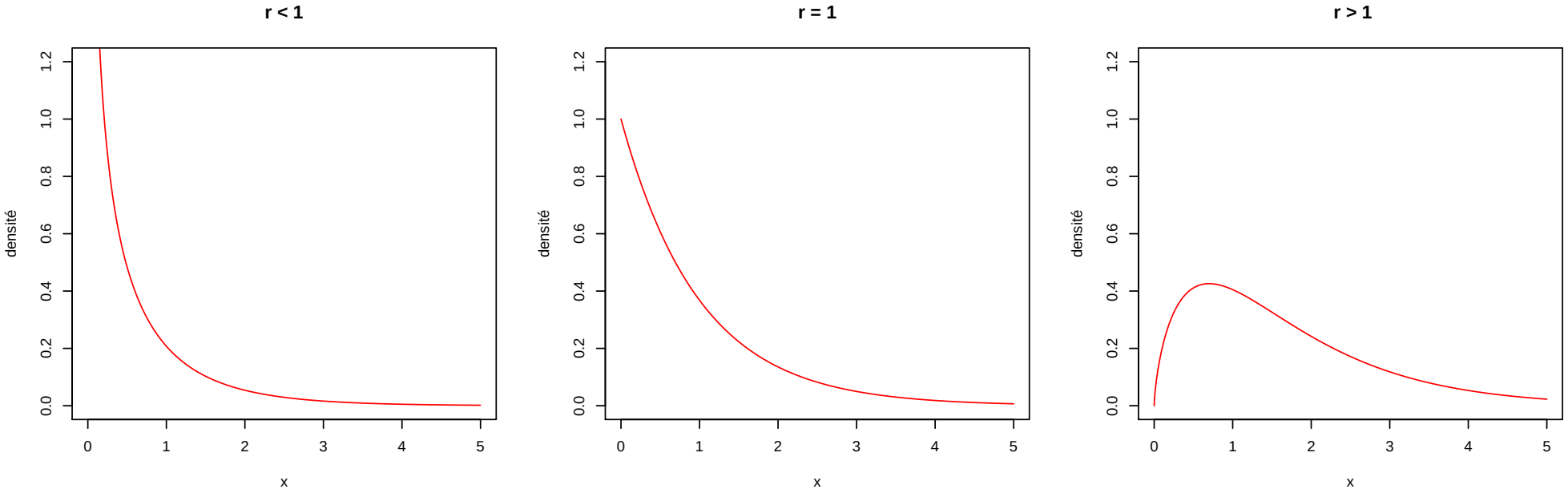
Compter les papillons

Parenthèse : la loi $\Gamma(r,\theta)$

C'est une loi à densité. La densité de $\Gamma(r,1)$ est (pour $r>0$)

$$\frac{1}{(r-1)!}x^{r-1}e^{-x} \quad (x>0)$$

(r n'est pas forcément entier ; on peut étendre la fonction factorielle aux réels >-1)



Compter les papillons

Parenthèse : la loi $\Gamma(r,\theta)$

C'est une loi à densité. La densité de $\Gamma(r,1)$ est (pour $r>0$)

$$\frac{1}{(r-1)!}x^{r-1}e^{-x} \quad (x>0)$$

- Si $X \sim \Gamma(r,1)$, $Y=(\theta X) \sim \Gamma(r,\theta)$ (paramètre d'échelle $\theta>0$)
- $E(X)=\mathrm{Var}(X)=r$
- $E(Y)=r\theta$, $\mathrm{Var}(Y)=r\theta^2$.

Compter les papillons

Combien de papillons d'une même espèce ?

Dans ce modèle, la probabilité d'observer $X = k$ papillons dans une espèce quelconque est

$$\mathbb{P}(X = k) = \int_\lambda \mathbb{P}_\lambda(X = k) f(\lambda) d\lambda$$

où $\mathbb{P}_\lambda(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$ (loi de Poisson) et $f(\lambda)$ est la densité de la loi $\Gamma(r, \theta)$.

(C'est simplement la formule des probabilités totales !)

$$\boxed{\mathbb{P}(X = k) = \frac{1}{(1+\theta)^r} \times \frac{(r+k-1)!}{(r-1)!k!} \times \left(\frac{\theta}{1+\theta}\right)^k}$$

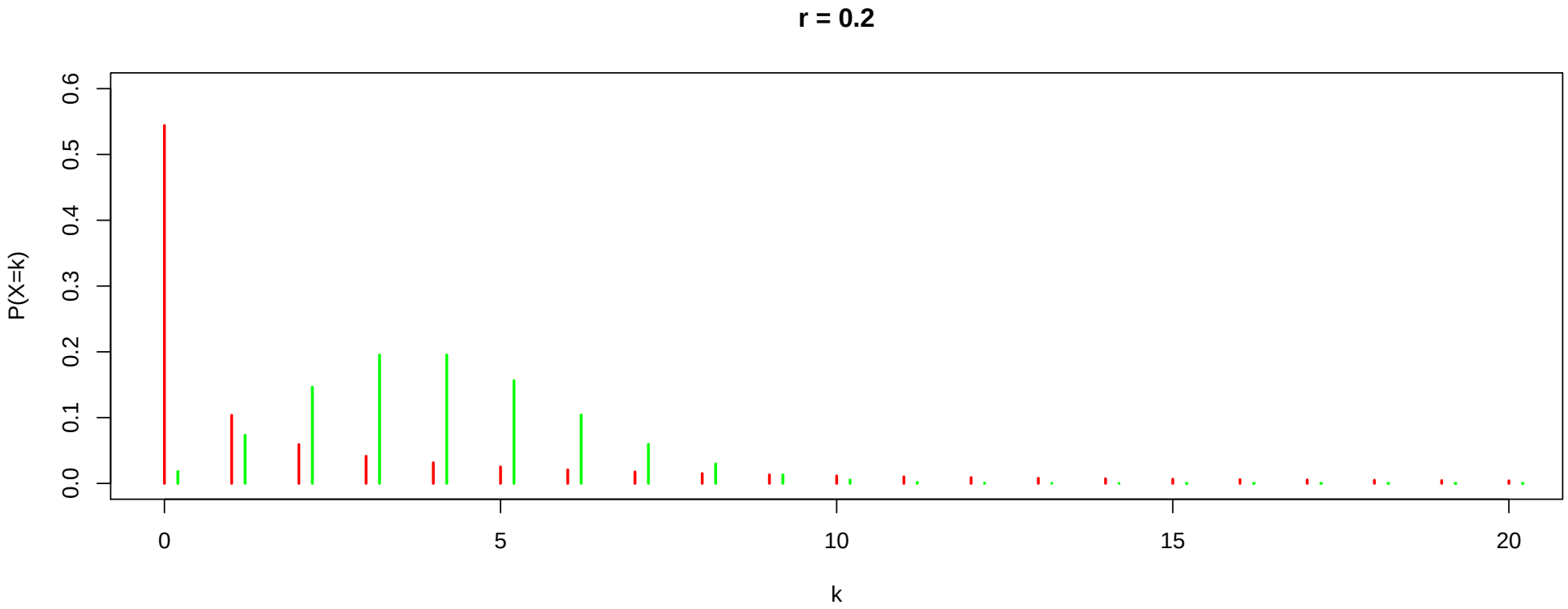
(Et ça c'est la « loi binomiale négative » !)

Compter les papillons

La loi binomiale négative

Son espérance est $E(X) = r\theta$ et sa variance $\mathrm{Var}(X) = r\theta(1+\theta)$.

On peut interpréter r comme un paramètre de sur-dispersion : *à espérance* $r\theta$ *fixée*, plus r est grand, plus la loi ressemble à une loi de Poisson – toutes les espèces ont à peu près le même paramètre λ ; les petites valeurs de r correspondent au contraire à une hétérogénéité importante.

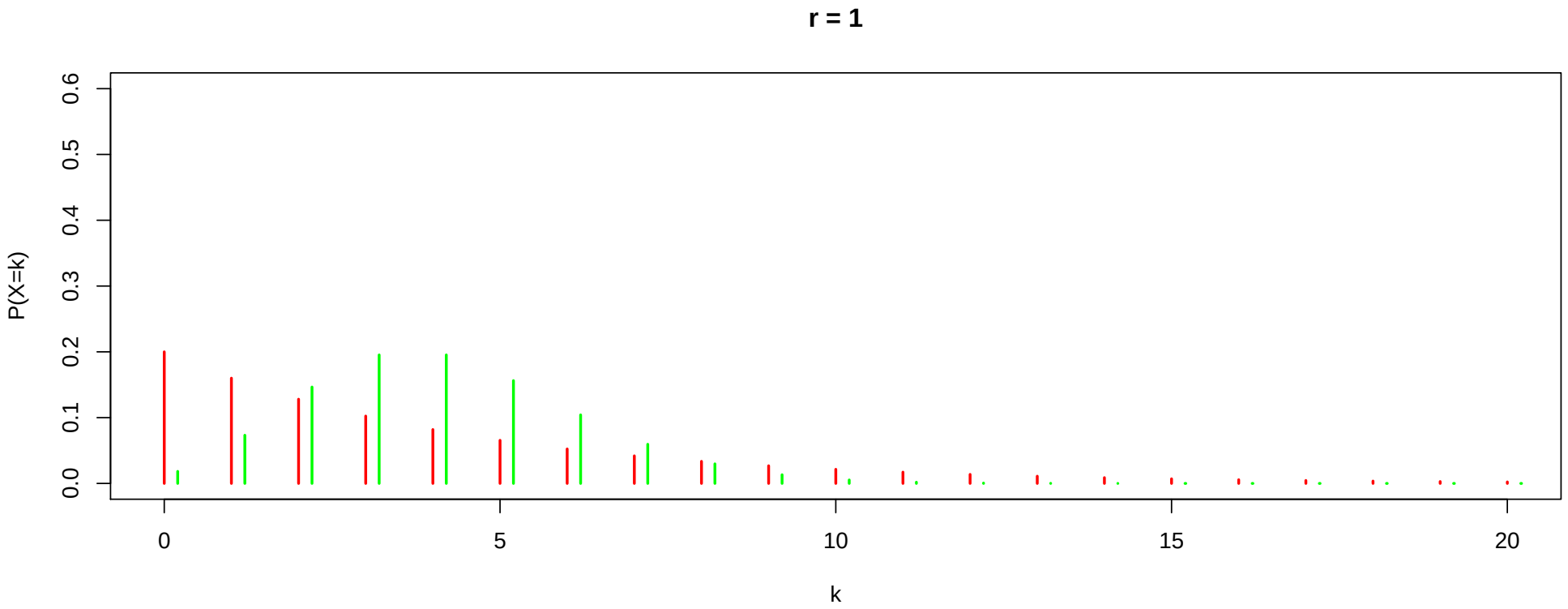


Compter les papillons

La loi binomiale négative

Son espérance est $E(X) = r\theta$ et sa variance $\mathrm{Var}(X) = r\theta(1+\theta)$.

On peut interpréter r comme un paramètre de sur-dispersion : *à espérance* $r\theta$ *fixée*, plus r est grand, plus la loi ressemble à une loi de Poisson – toutes les espèces ont à peu près le même paramètre λ ; les petites valeurs de r correspondent au contraire à une hétérogénéité importante.

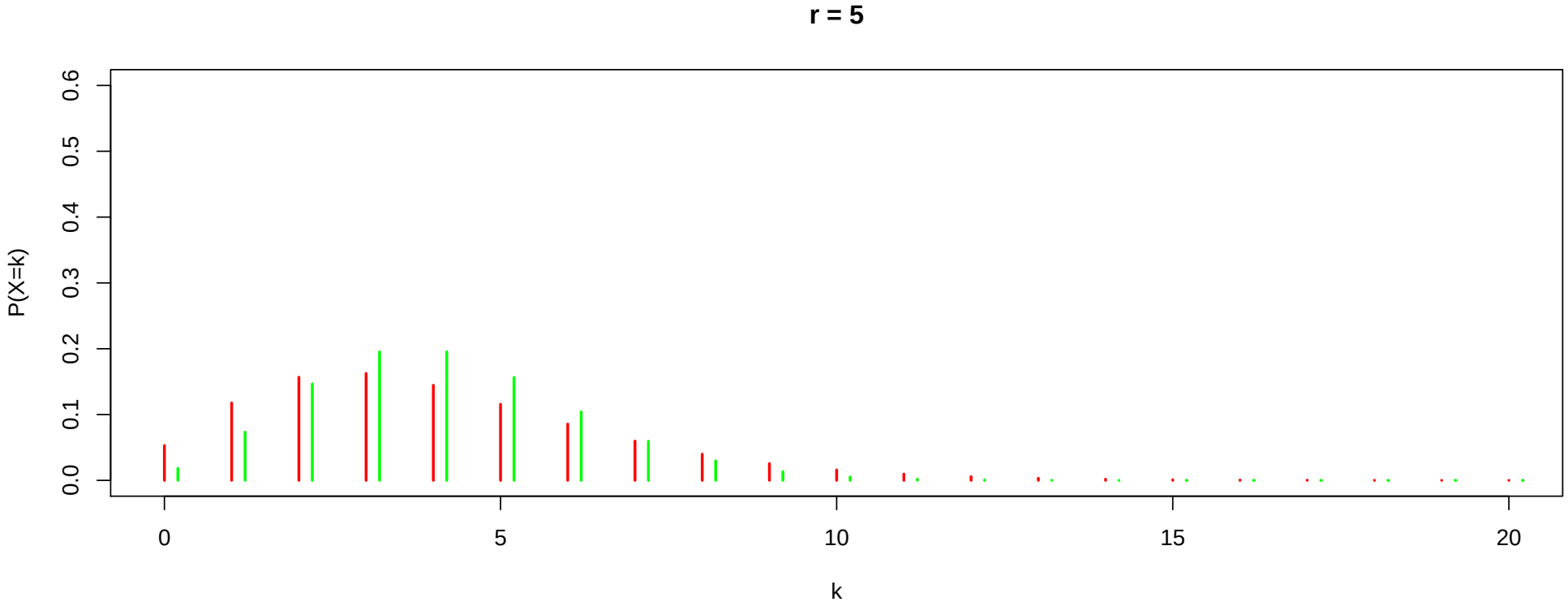


Compter les papillons

La loi binomiale négative

Son espérance est $E(X) = r\theta$ et sa variance $\mathrm{Var}(X) = r\theta(1+\theta)$.

On peut interpréter r comme un paramètre de sur-dispersion : *à espérance* $r\theta$ *fixée*, plus r est grand, plus la loi ressemble à une loi de Poisson – toutes les espèces ont à peu près le même paramètre λ ; les petites valeurs de r correspondent au contraire à une hétérogénéité importante.

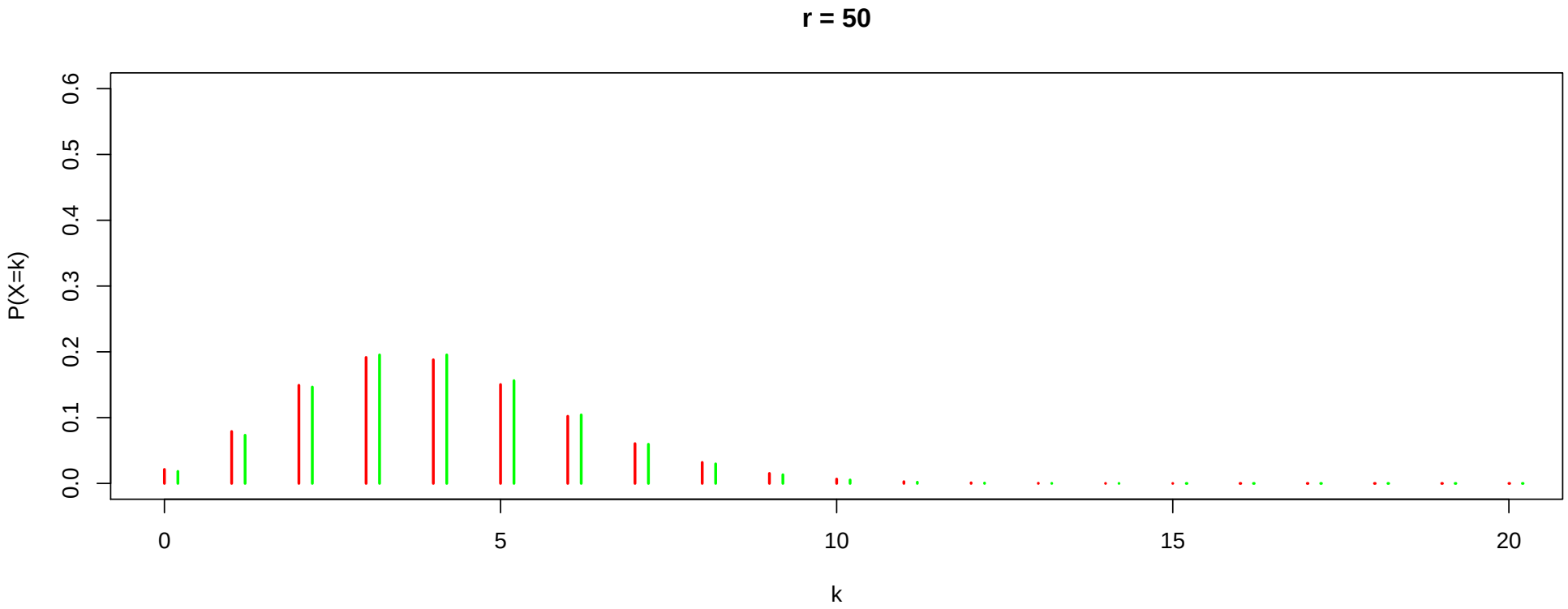


Compter les papillons

La loi binomiale négative

Son espérance est $E(X) = r\theta$ et sa variance $\mathrm{Var}(X) = r\theta(1+\theta)$.

On peut interpréter r comme un paramètre de sur-dispersion : *à espérance* $r\theta$ *fixée*, plus r est grand, plus la loi ressemble à une loi de Poisson – toutes les espèces ont à peu près le même paramètre λ ; les petites valeurs de r correspondent au contraire à une hétérogénéité importante.



Compter les papillons

Revenons à nos moutons : compter les espèces non vues

Supposons qu'on connaisse θ et r. Si s est le nombre (inconnu) total d'espèces présentes dans le milieu, on a

$$E(S_k) = s \times \mathbb{P}(X = k)$$

et l'espérance du nombre total d'espèces observées est

$$E(S_1 + S_2 + \cdots) = s \times \mathbb{P}(X > 0) = s \times (1 - (1+\theta)^{-r})$$

On peut estimer s par

$$\boxed{\widehat{s} = \frac{\sum_{k \geq 1} S_k}{1 - (1+\theta)^{-r}}.}$$

Compter les papillons

Revenons à nos moutons : prédire les résultats d'une nouvelle campagne d'observation

Si on fait une nouvelle campagne d'observation de durée t, le nombre attendu d'espèces non vues auparavant est

$$\begin{aligned}\Delta(t) &= s\int_\lambda \mathbb{P}_\lambda(X=0)\mathbb{P}_{\lambda t}(X>0)f(\lambda)d\lambda \\ &= s\int_\lambda e^{-\lambda}(1-e^{-\lambda t})f(\lambda)d\lambda\end{aligned}$$

On trouve

$$\Delta(t) = s(1+\theta)^{-r}\left(1-\left(1+\frac{\theta}{1+\theta}t\right)^{-r}\right)$$

De façon rassurante, quand t tend vers $+\infty$, $\Delta(t)$ tend vers $E(S_0)$ (le nombre d'espèces non vues).

Compter les papillons

Estimer les paramètres θ et r

Avant de mettre en branle la machinerie de la vraisemblance, remarquons qu'on n'observe pas S_0.

☞ Probabilité d'observer $X = k$ sachant que $X > 0$:

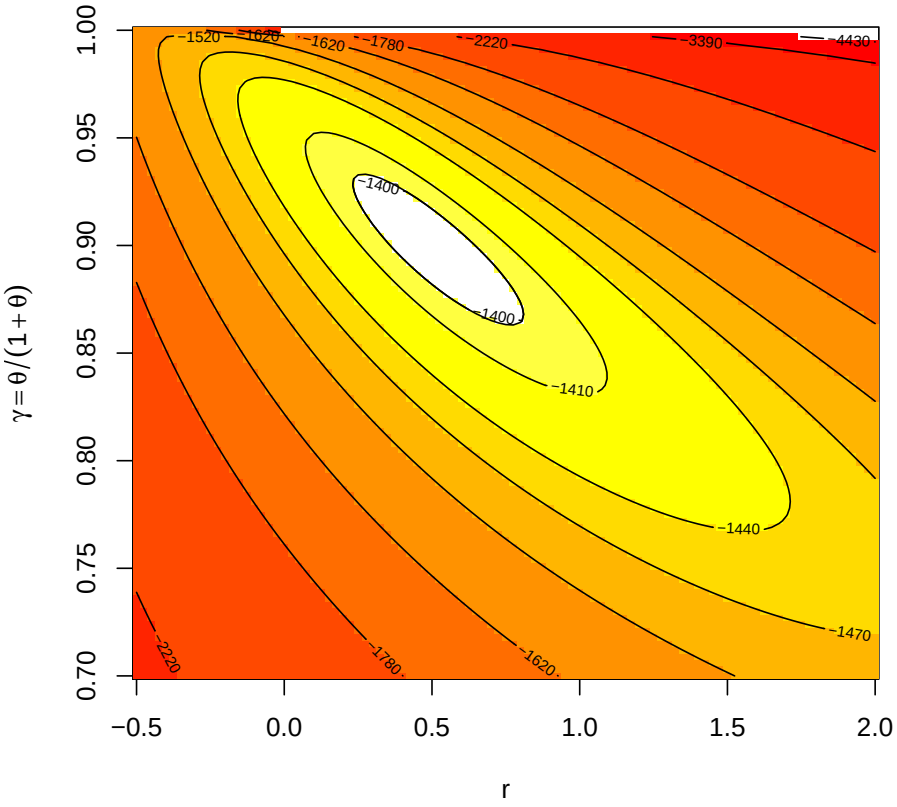
$$\begin{aligned}\mathbb{P}(X = k | X > 0) &= \frac{\mathbb{P}(X = k)}{1 - \mathbb{P}(X = 0)} \\ &\propto \frac{(r + k - 1)!}{k!} \left(\frac{\theta}{1 + \theta} \right)^k\end{aligned}$$

Avec $k > 0$, cette quantité a un sens pour $r > -1$ (et non plus seulement pour $r > 0$) (à condition d'adapter la constante de normalisation).

Compter les papillons

Estimer les paramètres θ et r : la vraisemblance

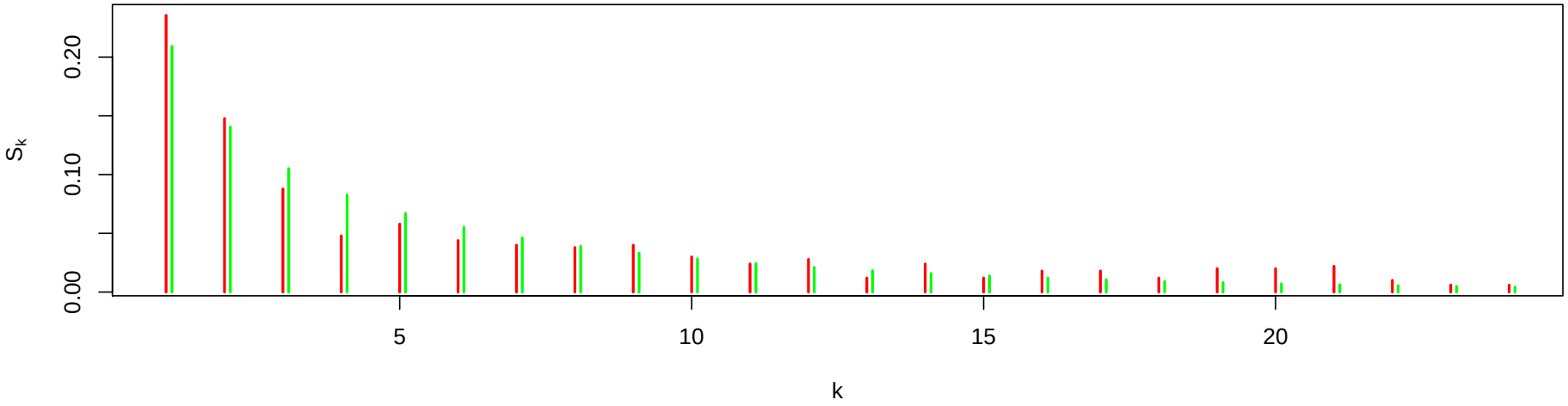
$$L(r,\theta) = \prod_k \mathbb{P}(X = k | X > 0)^{S_k}$$



Compter les papillons

Estimer les paramètres θ et r

Le maximum de vraisemblance est pour $r = 0.49$ et $\gamma = 0.9$.



On en déduit une estimation du nombre d'espèces : $\widehat{s} = 738$.

Compter les papillons

Quel crédit accorder à ce résultat ?

- Le résultat dépend d'hypothèses discutables
- Le problème le plus flagrant est le choix d'une loi Γ dans laquelle les fréquences relatives des espèces seraient tirées. Ce choix n'est dicté que par un souci de « commodité »...
- Que ferait-on si le maximum de vraisemblance était pour $r \leq 0$? ! (nombre infini d'espèces)
- Remarque : Fisher prend $r = 0$ et $\mathbb{P}(X = k | X > 0) = \frac{-1}{\log(1-\gamma)} \frac{\gamma^k}{k}$.

Pointe sèche de Paul-César Helleu

Le vocabulaire de Proust

De *a* à *zut*

- Dans la *Recherche*, il y a 1 229 224 mots en tout, et 41 220 mots distincts ; « a » apparaît 4 199 fois, « zut » 7 fois.
- On distingue ici toutes les formes conjuguées d'un même verbe, etc. (comme Efron le fait pour Shakespeare, ce qui n'est pas une bonne excuse)
- On considère les 100 000 premiers mots ; que peut-on prédire des 10 000 mots suivants ?
  - Nombre de nouveaux mots ?
  - Nombre de mots ayant déjà été utilisés une fois ?

Le vocabulaire de Proust

Les 100 000 premiers mots

Il y a $S_1 = 6398$ mots qui n'apparaîssent qu'une fois, S_2 1860 qui apparaissent 2 fois... un mot qui apparaît 4482 fois, « de ».
On ne montre ci-dessous que les premières valeurs de S_k.

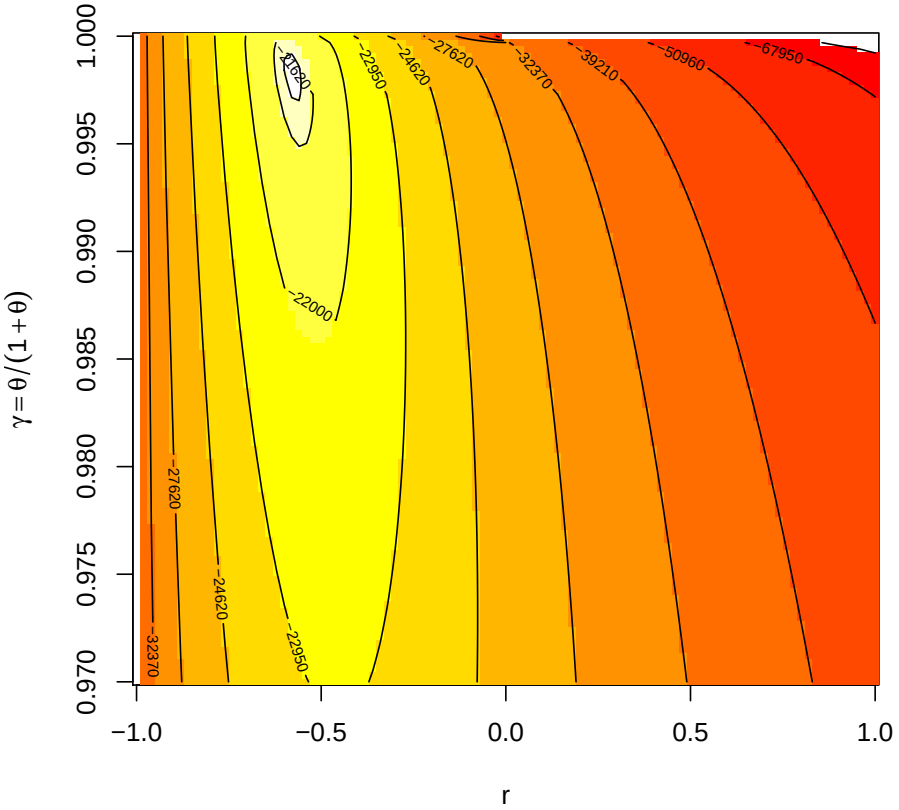
k	1	2	3	4	5	6	7	8	9	10
S_k	6398	1860	867	483	345	242	156	140	121	71
k	11	12	13	14	15	16	17	18	19	20
S_k	68	48	45	43	39	31	37	34	20	26

Peut-on utiliser notre modèle ? (Chaque mot est un papillon)

Le vocabulaire de Proust

La vraisemblance

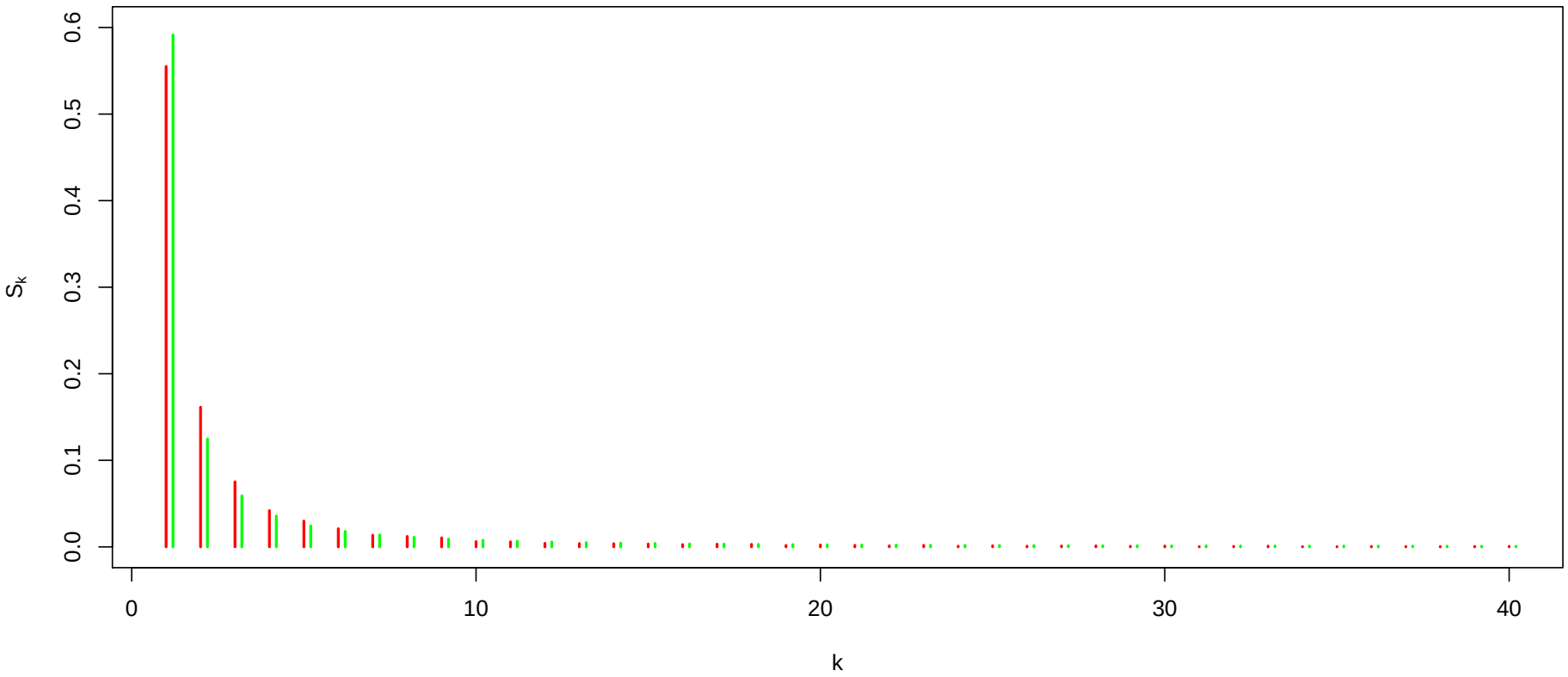
Le maximum de vraisemblance est pour $r < 0$...



...notre modèle n'a plus de sens.

Le vocabulaire de Proust

De plus l'adéquation n'est pas excellente...



Le vocabulaire de Proust

Version non paramétrique

On garde la même idée, mais on n'impose plus les λ pris dans une loi Γ.

- Les mots arrivent au fil de la plume selon un processus de Poisson
- Les λ de ce processus de Poisson sont pris au hasard dans une loi inconnue

Sans spécifier le moins du monde une famille de lois pour les λ, on va pouvoir écrire des prédicteurs.

Le vocabulaire de Proust

Nombre de nouveaux mots

On considère les 10 000 mots suivants. Combien apparaissent de mots qu'on n'avait pas observés parmi les 100 000 premiers ?

$$\begin{aligned}\Delta(t) &= s\int_\lambda \mathbb{P}_\lambda(X=0)\mathbb{P}_{\lambda t}(X>0)f(\lambda)d\lambda \\ &= s\int_\lambda e^{-\lambda}(1-e^{-\lambda t})f(\lambda)d\lambda\end{aligned}$$

En écrivant

$$1-e^{-\lambda t} = \lambda t - \frac{1}{2!}\lambda^2 t^2 + \frac{1}{3!}\lambda^3 t^3 - \cdots$$

et

$$E(S_k) = s\int_\lambda \frac{1}{k!}t^k e^{-\lambda}d\lambda,$$

on obtient

$$\Delta(t) = E(S_1)t - E(S_2)t^2 + E(S_3)t^3 - \cdots$$

D'où l'estimateur suivant :

$$\boxed{\widehat{\Delta}(t) = S_1 t - S_2 t^2 + S_3 t^3 - \cdots}$$

Avec $t = 0{,}1$ on prédit la présence de $\Delta(t) = 622$ nouveaux mots ; on en observe en fait 565 (bonne prédiction !)

Cet estimateur n'est utilisable que pour de petites valeurs de t : sa limite en $t \to +\infty$ n'a aucun sens.

Le vocabulaire de Proust

Fréquence des mots rares : estimateur de Good-Turing

Une estimation naïve de la fréquence d'un mot rencontré une fois sur les 100 000 premiers mots est 1 pour 100 000.
Si la loi a priori de λ est de « densité » $f(\lambda)$, la loi a posteriori est

$$f(\lambda|X=k) = \frac{1}{\kappa}\mathbb{P}_\lambda(X=k)f(\lambda)$$

avec une constante de normalisation

$$\kappa = \int_\lambda \mathbb{P}_\lambda(X=k)f(\lambda) = E(S_k)/s$$

Le vocabulaire de Proust

Fréquence des mots rares : estimateur de Good-Turing

L'estimateur bayésien de λ est

$$\begin{aligned}
E(\lambda|X=k) &= \int_\lambda \lambda \times f(\lambda|X=k)d\lambda \\
&= \frac{1}{\kappa}\int_\lambda \lambda \times \frac{\lambda^k}{k!}e^{-\lambda}f(\lambda)d\lambda \\
&= \frac{k+1}{\kappa}\int_\lambda \frac{\lambda^{k+1}}{(k+1)!}e^{-\lambda}f(\lambda)d\lambda \\
&= \frac{k+1}{\kappa}E(S_{k+1})/s = (k+1)\frac{E(S_{k+1})}{E(S_k)}
\end{aligned}$$

On a donc l'estimateur suivant

$$\boxed{\widehat{E}(\lambda|k) = (k+1)\frac{S_{k+1}}{S_k}}$$

Le vocabulaire de Proust

Fréquence des mots rares : estimateur de Good-Turing

Avec $k = 1$ on estime la fréquence des mots qui sont apparus une fois parmi les 100 000 premiers par

$$2 \times \frac{S_2}{S_1} = 2 \times \frac{1860}{6398} = 0,58$$

apparitions tous les 100 000 mots.

On prédit donc que chacun de ces 6398 mots va apparaître, en espérance, 0,058 fois dans les 10 000 mots suivants.

On doit donc compter environ $6398 \times 0,058 = 371$ apparitions de mots présents une seuls fois dans les 100 000 premiers mots : il y en a en fait 317.

(Si on leur avait assigné une fréquence de 1 pour 100 000 on attendait de les voir apparaître 640 fois !)

Le vocabulaire de Proust

What else ?

- On peut utiliser ces estimateurs pour construire un test pour vérifier si un morceau de texte appartient à *la Recherche* ou non... (Efron et Thisted le font pour Shakespeare)
- Ce modèle est-il utilisable pour faire du clustering ?
- Peut-on introduire un prior plus complexe, construit à partir du « reste de la littérature » ? (par exemple on sait que « de », « le », « la », « et »... sont fréquents)

Bibliographie partielle

Fisher et al 1943, The relation between the number of species and the number of individuals in a random sample of an animal population

Good, 1953. The population frequencies of species and the estimation of population parameters.

Good and Toulmin, 1956. The number of new species, and the increase in population coverage, when a sample is increased.

Efron et Thisted, 1976. Estimating the number of unseen species : how many words did Shakespeare know ?

Thisted et Efron, 1979. Did Shakespeare write a newly-discovered poem ?

pou pou pidou