T.D. Inférence par maximum de vraisemblance sur des lois discrètes
–
corrigé

Philippe Veber

27 février 2020

Exercice 1. Soit une population dans laquelle on s'intéresse à la prévalence p d'une pathologie. Pour estimer p, on constitue un échantillon de 5 individus choisis uniformément et indépendamment. On fait passer à chaque individu un test supposé fiable pour détecter la pathologie, qui s'avère positif pour les individus 1, 2 et 5, et négatif pour les autres. Proposez un modèle des données obtenues et une méthode pour estimer p.

On note M_i le résultat du test (0 pour négatif, 1 pour positif) pour la i^{e} personne. Il s'agit de variables binaires, indépendantes puisqu'elles correspondent à un tirage aléatoire uniforme dans une population. On a donc le modèle suivant :

$$M_i \sim \text{Bern}(p) \text{ pour } i \in \{1, \dots, 5\}$$

Pour estimer p, on procède par maximum de vraisemblance. La fonction de vraisemblance s'écrit pour ce modèle :

$$\begin{aligned} L(p \mid M_1, \dots, M_5) &= \mathbb{P}[M_1, \dots, M_5 \mid p] \\ &= \mathbb{P}[M_1 \mid p] \dots \mathbb{P}[M_5 \mid p] \text{ car les } M_i \text{ sont indépendants conditionnellement à } p \\ &= pp(1-p)(1-p)p \\ &= p^3(1-p)^2 \end{aligned}$$

Pour simplifier les calculs suivants, on calcule maintenant la log-vraisemblance, qui est maximale si et seulement si la vraisemblance l'est, car la fonction log est strictement croissante. Attention néanmoins, la log-vraisemblance n'est définie que si la vraisemblance est non-nulle, c'est-à-dire ici $p \neq 0$ et $p \neq 1$. Puisque nous cherchons un maximum de la vraisemblance et que celle-ci est positive, nous pouvons sans danger exclure ces valeurs pour p. On obtient ainsi :

$$\log L(p; M_1, \dots, M_5) = 3 \log p + 2 \log(1-p)$$

Cette fonction est une somme de deux fonctions concaves, elle est donc concave également.

Par conséquent, tout point qui annule la dérivée est un maximum. En posant $f = \log L$,

$$f'(p) = \frac{3}{p} - \frac{2}{1-p}$$

On résout maintenant $f'(p) = 0$ pour obtenir :

$$\begin{aligned} \frac{3}{p} &= \frac{2}{1-p} \\ 3(1-p) &= 2p \\ p &= \frac{3}{5} \end{aligned}$$

L'estimation en maximum de vraisemblance $\hat{p}$ de p est donc $\frac{3}{5}$.

La loi de Bernoulli s'exprime avec une expression conditionnelle (si ... alors ... sinon ...), ce qui n'était pas un souci dans l'exercice précédent puisque l'on savait pour chaque variable sa valeur observée. Pour les cas $M_i = 1$, nous avons mis un terme p et pour les cas $M_i = 0$ un terme $1 - p$. Il est possible d'exprimer ce choix avec une simple expression arithmétique, en remarquant que pour tout réel x, $x^0 = 1$ et $x^1 = x$. Ainsi pour une variable $X \sim \text{Bern}(p)$, sa loi de probabilité peut s'écrire :

$$\mathbb{P}[X = x \mid p] = p^x (1-p)^{(1-x)} \text{ pour } x \in \{0, 1\}$$

Cette écriture est mise à profit dans l'exercice suivant.

Exercice 2. Soit un échantillon de variables indépendantes $X_1, \ldots, X_n$ distribuées selon $\text{Bern}(p)$. Déterminez l'estimateur en maximum de vraisemblance de p.

Pour estimer p, on procède par maximum de vraisemblance. La fonction de vraisemblance s'écrit pour ce modèle :

$$\begin{aligned} L(p \mid X_1, \ldots, X_n) &= \mathbb{P}[X_1, \ldots, X_n \mid p] \\ &= \prod_{i=1}^{n} \mathbb{P}[X_i \mid p] \text{ car les } M_i \text{ sont indépendants conditionnellement à } p \\ &= \prod_{i=1}^{n} p^{X_i} (1-p)^{(1-X_i)} \end{aligned}$$

c'est-à-dire sous la forme d'un produit où le terme p apparaît autant de fois qu'il y a d'individus malades dans l'échantillon, et $1 - p$ autant de fois qu'il y a d'individus sains. Appelons k le nombre d'individididus malades, on a :

$$k = \sum_{i=1}^{n} X_i$$

et la vraisemblance peut alors s'écrire :

$$L(p \mid X_1, \ldots, X_n) = p^k (1-p)^{n-k}$$

On en déduit l'expression suivante pour la log-vraisemblance, pour $p \in]0;1[$:

$$\log L(p \mid X_1, \ldots, X_n) = k \log p + (n-k) \log(1-p)$$

La fonction $f = \log L$ est une somme de deux fonctions concaves, elle est donc elle-même concave. Son maximum est unique s'il existe et il suffit d'annuler la dérivée $f'(x) = \frac{k}{p} - \frac{n-k}{1-p}$ pour le trouver. Cela donne :

$$\hat{p} = \frac{k}{n}$$

Notons bien que k, malgré la notation choisie, est bien une variable aléatoire, puisqu'elle est une somme de variables aléatoires, et par conséquent $\hat{p}$ aussi.

Exercice 3. On réalise un test de toxicité d'un produit chimique, en exposant des crustacés de type Daphnie à une concentration contrôlée du produit.

1. On prépare un aquarium avec n individus, et après 12h d'exposition on compte le nombre k de survivants. Explicitez un modèle très simple de l'expérience et dérivez-en une estimation du taux de survie des crustacés. Explicitez les deux hypothèses qui permettent de considérer ce modèle.

2. On prépare maintenant deux aquariums de n_1 et n_2 individus respectivement, et l'on obtient à l'issue de l'expérience respectivement k_1 et k_2 survivants. À nouveau, explicitez un modèle simple de cette expérience et dérivez-en une estimation du taux de survie. Sous ce modèle, quel est l'intérêt d'avoir préparé deux aquariums ?

1. On pose deux hypothèses simplificatrices : d'une part tous les individus ont la même sensibilité au produit chimique, d'autre part la survie d'un individu n'a pas d'impact sur la survie des autres. On peut alors représenter la survie des individus par une collection de v.a. distribuées selon une loi de Bernoulli de même paramètre p indépendantes. Le nombre d'individus ayant survécu est alors distribué selon une loi binomiale de paramètres n et p. Notre modèle est ainsi :

p : taux de survie d'un individu à une exposition de 12h au produit chimique

n : nombre d'individus initialement présents dans l'aquarium

k : nombre de survivants observé à l'issue de l'expérience

$k \sim \text{Binom}(n, p)$

La vraisemblance des données sous ce modèle s'écrit :

$$L(p \mid k) = \mathbb{P}[k \mid p] = \binom{n}{k} p^k (1-p)^{(n-k)}$$

et la log-vraisemblance (en excluant les valeurs 0 et 1 pour p) :

$$\log L(p \mid k) = \log \binom{n}{k} + k \log p + (n-k) \log(1-p)$$

On retrouve ici à une constante additive près la vraisemblance de l'exercice 2. Il ne faut pas en être surpris puisque le calcul avait alors fait apparaître que l'estimation en maximum de vraisemblance dépendait seulement du nombre de succès, c'est-à-dire la grandeur qui suit précisément une loi binomiale. Au final, on tombe sur la même estimation de p :

$$\hat{p} = \frac{k}{n}$$

2. En posant que ce qui se passe dans un aquarium n'influence pas ce qui se passe dans l'autre, on peut proposer le modèle suivant :

p : taux de survie d'un individu à une exposition de 12h au produit chimique
n : nombre d'individus initialement présents dans l'aquarium
k_i : nombre de survivants observé à l'issue de l'expérience dans l'aquarium $i \in \{1,2\}$
$k_i \sim \text{Binom}(n,p)$ pour $i \in \{1,2\}$

La vraisemblance sous ce modèle s'écrit :

$$\begin{aligned} L(p \mid k_1,k_2) &= \mathbb{P}[k_1,k_2 \mid p] \\ &= \mathbb{P}[k_1 \mid p]\mathbb{P}[k_2 \mid p] \quad \text{(indépendance entre aquariums conditionnellement à } p) \\ &= \binom{n_1}{k_1} p^{k_1}(1-p)^{(n_1-k_1)} \binom{n_2}{k_2} p^{k_2}(1-p)^{(n_2-k_2)} \\ &= \binom{n_1}{k_1}\binom{n_2}{k_2} p^{k_1+k_2}(1-p)^{(n_1+n_2-k_1-k_2)} \end{aligned}$$

Avec le même raisonnement que précédemment, on arrive à l'estimation :

$$\hat{p} = \frac{k_1+k_2}{n_1+n_2}$$

On constate que l'on arrive à la même estimation que si l'on avait eu un seul aquarium contenant n_1+n_2 individus et observé k_1+k_2 survivants. Sous ce modèle, il n'y a donc pas de valeur ajoutée à préparer des réplicats biologiques sous la forme de plusieurs aquariums.

Exercice 4. On considère un alignement nucléique de n séquences et on s'intéresse à une colonne particulière de cet alignement. Le nucléotide trouvé à cette position dans la séquence i est supposé suivre la distribution :

$$N_i \sim \text{Cat}(\mathcal{N}, p)$$

où $\mathcal{N}$ est l'ensemble des nucléotides et p la probabilité de chaque nucléotide à cette position.

1. D'après ce modèle, diriez-vous que les séquences sont indépendantes les unes des autres, ou non-indépendantes ?

2. Écrivez la fonction de vraisemblance en fonction des variables p_A, p_C et p_G pour $n = 5$ et la colonne d'alignement $C = (\texttt{A T A G C})$

3. En examinant la réponse à la question précédente, expliquez pourquoi il suffit de décrire la colonne d'alignement par les comptages k_A, k_C et k_G des nucléotides A, C et G pour calculer la vraisemblance de la colonne.

4. Déterminez l'estimation en maximum de vraisemblance pour le cas particulier de la question 2.

5. Écrivez la fonction de vraisemblance pour une colonne d'alignement et une valeur de n quelconque en fonction de p_A, p_C, p_G et p_T et des comptages k_A, k_C, k_G et k_T.

6. (difficile) Déduire de la question précédente l'estimation en maximum de vraisemblance de p.

1. Le modèle indique que les variables N_i sont tirées indépendamment selon une même distribution.

2.

$$\begin{aligned} L(p \mid N_1, \ldots, N_5) &= \mathbb{P}[N_1, \ldots, N_5 | p] \\ &= \mathbb{P}[N_1|p]\mathbb{P}[N_2|p] \ldots \mathbb{P}[N_5|p] \quad \text{(indépendance conditionnellement à } p\text{)} \\ &= p_A p_T p_A p_C p_C \\ &= p_A(1 - p_A - p_C - p_G) p_A p_G p_C \\ &= {p_A}^2(1 - p_A - p_C - p_G) p_G p_C \end{aligned}$$

3. L'expression de la vraisemblance est constituée de 4 termes p_A, p_C, p_G et $1 - p_A - p_C - p_G$ apparaissant respectivement avec les puissances k_A, k_C, k_G et $k_G = n - k_A - k_C - k_G$. Ainsi, seuls les comptages interviennent, pas l'assignation exacte des nucléotides à chaque séquence.

4. On pose $f = \log L$ la log-vraisemblance des paramètres qui admet les mêmes maxima que L. Elle s'écrit :

$$f(p_A, p_C, p_G) = 2 \log p_A + \log p_C + \log p_G + \log(1 - p_A - p_C - p_G)$$

Il s'agit d'une fonction concave, il suffit donc de déterminer un point d'annulation du gradient pour trouver un maximum. On calcule donc les dérivées partielles de f :

$$\begin{aligned} \frac{\partial f}{\partial p_A} &= \frac{2}{p_A} - \frac{1}{1 - p_A - p_C - p_G} \\ \frac{\partial f}{\partial p_C} &= \frac{1}{p_C} - \frac{1}{1 - p_A - p_C - p_G} \\ \frac{\partial f}{\partial p_G} &= \frac{1}{p_C} - \frac{1}{1 - p_A - p_C - p_G} \end{aligned}$$

On en tire :

$$\begin{aligned} \frac{\partial f}{\partial p_A} = 0 &\Rightarrow p_A = 2(1 - p_A - p_C - p_G) \\ \frac{\partial f}{\partial p_C} = 0 &\Rightarrow p_C = 1 - p_A - p_C - p_G \\ \frac{\partial f}{\partial p_G} = 0 &\Rightarrow p_G = 1 - p_A - p_C - p_G \end{aligned}$$

D'où :

$$p_G = p_C$$
$$p_C = \frac{1 - p_A}{3}$$
$$3p_A = 2(1 - \frac{2}{3}(1 - p_A))$$

On en déduit enfin $p_A = \frac{2}{5}$, $p_C = \frac{1}{5}$, $p_G = \frac{1}{5}$ et $p_T = \frac{1}{5}$.

5.

$$\begin{aligned} L(p; N_1, \ldots, N_n) &= \mathbb{P}[N_1, \ldots, N_n | p] \\ &= \mathbb{P}[N_1|p]\mathbb{P}[N_2|p] \ldots \mathbb{P}[N_n|p] \quad \text{(indépendance conditionnellement à } p\text{)} \\ &= \prod_{i=1}^{n} p_{N_i} \\ &= \prod_{x \in \mathcal{N}} p_x{}^{\sum_{i=1}^{n} \mathbb{1}[N_i = x]} \\ &= {p_A}^{k_A} {p_C}^{k_C} {p_G}^{k_G} {p_T}^{k_T} \end{aligned}$$

6. Pour commencer, remarquons que si pour un nucléotide i, on a $k_i = 0$, alors pour toute valeur de p, on peut augmenter la vraisemblance en mettant p_i à o. En effet, les autres paramètres doivent alors être augmentés pour satisfaire la contrainte $\sum_i p_i = 1$ et cela ne peut pas décroître la vraisemblance (calculer la dérivée pour s'en convaincre). Donc sans perte de généralité, on peut poser que pour tout k_i nul, le p_i correspondant vaut o et restreindre ce qui suit au nucléotides de comptage non nul. Pour les nucléotides de comptage non nul, si le paramètre p_i correspondant vaut o alors la vraisemblance vaut o également et ne saurait donc être maximale. On peut donc exclure cette possibilité et considérer la log-vraisemblance :

$$\log L(p; N_1, \ldots, N_n) = \sum_{i \in \mathcal{N}} k_i \log p_i$$

Pour s'affranchir de la contrainte entre les p_i on réécrit la fonction sous la forme :

$$f(x) = k_0 \log\left(1 - \sum_{i>0} p_i\right) + \sum_{i>0} \log p_i$$

Ainsi toutes les variables de f sont indépendantes. Les k_i étant positifs, f est une somme de fonctions concaves, elle est donc elle-même concave et admet au plus un maximum. Si ce maximum existe il doit annuler le gradient de f :

$$\frac{\partial f}{\partial p_i} = 0 \Rightarrow \frac{k_i}{p_i} = \frac{k_0}{1 - \sum_{j>0} p_j} = \frac{k_0}{p_0} \qquad \text{pour } i > 0$$

On en déduit que tous les rapports $\frac{k_i}{p_i}$ sont égaux ; notons λ cette valeur. On a alors :

$$\sum_i p_i = 1$$
$$\sum_i \frac{k_i}{\lambda} = 1$$
$$\lambda = \sum_i k_i$$

D'où l'on conclut :

$$p_i = \frac{k_i}{\sum_j k_j}$$

1 Loi de Poisson

La loi binomiale nous permet de modéliser un nombre d'évènements – appelés succès – lorsque l'on réalise un nombre fixé de tests (ou épreuves). Considérons maintenant une autre situation, celle où l'on observe des évènements arriver dans un intervalle de temps fixé.

Quand ces évènements surviennent (1) avec un taux constant dans le temps, (2) de manière indépendante, (3) et de telle sorte que dans un laps de temps très court, la probabilité de voir deux évènements devient négligeable, alors le nombre total d'évènements survenant dans l'intervalle de temps considéré suit la loi de Poisson.

La loi de Poisson est une distribution discrète, définie sur l'ensemble des entiers naturels. Elle admet un paramètre λ réel positif correspondant au nombre moyen d'évènements observés sur l'intervalle considéré. On la note $\mathcal{P}(\lambda)$, et elle est définie par l'expression :

$$\forall k \in \mathbb{N},\ \mathbb{P}[X = k] = \frac{\lambda^k}{k!} e^{-\lambda}$$

La loi de Poisson est utile dans de nombreux contextes, voici quelques exemples d'utilisation :

— le nombre d'appels reçus dans un centre d'appel en une minute

— le nombre de divisions cellulaires observées sur une boîte de Pétri pendant une heure

— le nombre de mutations apparues sur un chromosome dans un temps donné

— le nombre de bactéries observées dans un volume donné d'une solution.

Le dernier exemple illustre que l'on retrouve de la même façon une loi de Poisson lorsque les évènements surviennent selon un taux par unité d'espace plutôt que de temps.

Exercice 5. Lors d'une étude sanitaire sur 100 lots de viande issus de différents établissements, on compte dans chaque échantillon récolté le nombre de parasites de type *Trichinella spiralis* présents. Après avoir chargé les données sous R, on calcule la moyenne et la variance des comptages trouvés :

```
> mean(counts)
[1] 1.02
> var(counts)
[1] 1.93899
```

On propose d'utiliser le modèle suivant :

n : nombre de lots étudiés
λ : nombre moyen de parasites par lot
k_i : nombre de parasites trouvés dans le lot i
$k_i \sim \mathcal{P}(\lambda)$

Selon vous, ce modèle est-il approprié ?

Selon le modèle proposé, le nombre de parasites trouvé dans chaque lot suit la même loi de Poisson. Par conséquent, on devrait trouver que la moyenne des comptages est égale à la variance. Or ce n'est pas le cas, l'écart ne pouvant raisonnablement être expliqué par la taille de l'échantillon. Pour s'en convaincre on pourrait par exemple faire une simulation, et estimer la fréquence avec laquelle on obtient une telle différence entre les deux en reprenant la même taille d'échantillon.

Exercice 6. On considère des échantillons d'eau de mer dans lesquels on compte le nombre de bactéries présentes.

1. Dans une eau dont la concentration en bactéries est de 0,2 unités par microlitre, quelle est la probabilité d'observer au moins 3 bactéries dans un échantillon de 10 µL ?

2. On dispose de n échantillons d'une même zone pour lesquels on a compté le nombre de bactéries présentes. Proposez une méthode pour estimer la concentration en bactéries dans la zone.

1. On modélise le nombre K de bactéries présentes dans un échantillon de 10 µL par une loi de Poisson de paramètre $\lambda = 0,2 \times 10 = 2$. La probabilité recherchée est :

$$\begin{aligned}\mathbb{P}[K \geq 3] &= 1 - \mathbb{P}[K \leq 2] \\ &= 1 - \mathbb{P}[K = 0] - \mathbb{P}[K = 1] - \mathbb{P}[K = 2] \\ &= 1 - e^{-\lambda}(1 + \lambda + \frac{\lambda^2}{2})\end{aligned}$$

Pour $\lambda = 2$, on a donc $\mathbb{P}[K \geq 3] \approx 0,323$

2. On supposera que tous les échantillons ont le même volume. Considérons le modèle suivant :

λ : nombre moyen de bactérie dans un échantillon
k_i : nombre de bactéries comptées dans le i^{e} échantillon
$k_i \sim \mathcal{P}(\lambda)$

On propose d'estimer λ à partir des comptages k_i par maximum de vraisemblance. La vrai-

semblance pour ce modèle s'écrit :

$$\begin{aligned} L(\lambda; k) &= \mathbb{P}[k_1, \ldots, k_n \mid \lambda] \\ &= \prod_{i=1}^{n} \mathbb{P}[k_i \mid \lambda] \text{ (indépendance des échantillons)} \\ &= \prod_{i=1}^{n} \frac{\lambda^{k_i}}{k_i!} e^{-\lambda} \\ &= \frac{\lambda^{\sum_{i=1}^{n} k_i}}{\prod_{i=1}^{n} k_i!} e^{-n\lambda} \end{aligned}$$

Calculons maintenant la log-vraisemblance $f = \log L$:

$$f(\lambda) = (\sum_{i=1}^{n} k_i) \log \lambda - n\lambda + \log \prod_{i=1}^{n} k_i!$$

La fonction f est dérivable et concave, on en trouve un maximum en annulant la dérivée :

$$f'(\lambda) = \frac{\sum_{i=1}^{n} k_i}{\lambda} - n$$

On déduit de $f'(\lambda) = 0$ l'estimateur :

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} k_i}{n}$$

Exercice 7. Nous avons rappelé que le nombre d'évènements survenant dans un intervalle de temps donné suivait la loi $\mathcal{P}(\lambda)$ si les évènements respectaient trois conditions (*cf supra*). Considérons maintenant la construction suivante : on découpe l'intervalle de temps en n portions égales, et pour chaque portion on définit une variable de Bernoulli de paramètre $p = \frac{\lambda}{n}$ indépendante [1], la variable indiquant si un évènement a lieu dans la portion considérée.

1. Quelle est la loi du nombre d'évènements dans cette construction ? Quelle est son espérance ?
2. Expliquez en quoi cette construction mime les trois conditions sur les évènements d'une loi de Poisson.
3. Reproduisez et commentez la figure 1.
4. Quelle relation cela suggère entre la loi de Poisson et la loi binomiale ?

1. La construction proposée peut être décrite formellement par

$$X_i \sim \text{Bern}(\frac{\lambda}{n}) \text{ pour } i \in \{1, \ldots, n\} K = \sum_{i=1}^{n} X_i$$

Les variables X_i étant des variables de Bernoulli indépendantes et identiquement distribuées,

1. il faut pour cela avoir *a minima* $n \geq \lambda$.

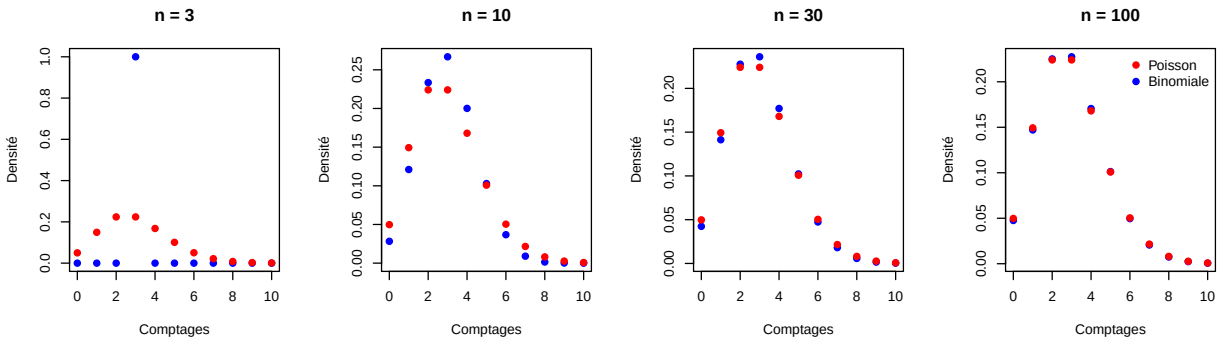

FIGURE 1 – Loi binomiale et loi de Poisson. Les graphes représentent en bleu la même distribution, la loi de Poisson de paramètre 3 ; en rouge, la distribution de lois binomiales de paramètre n et $\frac{3}{n}$.

K suit une loi binomiale de paramètres n et $\frac{\lambda}{n}$. On a ainsi :

$$\mathbb{E}[K] = \lambda$$

2. On revient sur les trois conditions :
   (a) tous les intervalles ont la même probabilité de voir une occurrence de l'évènement
   (b) les évènements surviennent indépendemment sur chaque intervalle
   (c) sur chaque intervalle, on voit au plus un évènement
3. Le code pour reproduire la figure

```
f <- function(n,l) {
  plot(0:10, dbinom(0:10, n, 3/n),
       pch=19, col="blue",
       xlab="Comptages",
       ylab="Densité",
       main=paste0("n = ",n))
  points(0:10,dpois(0:10, 3), col="red",pch=19)
  if(l) {
    legend("topright",
           legend=c("Poisson","Binomiale"),
           pch=19,col=c("red","blue"),
           bty="n")
  }
}
pdf("binom-conv.pdf",width=10,height=3)
par(mfrow=c(1,4))
f(3,F) ; f(10,F) ; f(30,F) ; f(100,T)
dev.off()
```

La figure présente pour quatre valeurs de n la comparaison entre une distribution $\mathcal{P}(\lambda)$ et $\mathrm{B}(n, \frac{\lambda}{n})$. On observe qu'en faisant croître n la distribution binomiale se rapproche de la distribution de Poisson.

4. La figure 1 suggère que la limite pour $n \to +\infty$ de $\mathrm{B}(n, \frac{\lambda}{n})$ est $\mathcal{P}(\lambda)$. Ce n'était pas demandé, mais en voici la preuve. Soit $X \sim \mathrm{B}(n, \frac{\lambda}{n})$. La distribution binomiale s'écrit :

$$\begin{aligned}\mathbb{P}[X = k] &= \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k (1-\frac{\lambda}{n})^{(n-k)} \\ &= \frac{\lambda^k}{k!}\underbrace{\left(1-\frac{\lambda}{n}\right)^n}_{(1)}\underbrace{\left(1-\frac{\lambda}{n}\right)^{-k}}_{(2)}\underbrace{\frac{n!}{(n-k)!n^k}}_{(3)}\end{aligned}$$

Le terme (3) converge vers 1 quand $n \to +\infty$, comme on le voit mieux sous cette forme :

$$\frac{n!}{(n-k)!n^k} = \prod_{i=1}^{k}\frac{n-i}{n}$$

Tous les termes du produit convergent vers 1, donc le produit converge vers 1. Le terme (2) converge vers 1 puisque $1 - \frac{\lambda}{n} \to 1$ quand $n \to +\infty$. Le terme (1) converge vers $e^{-\lambda}$. Pour s'en convaincre, on peut développer selon la formule du binôme de Newton :

$$\begin{aligned}(1-\frac{\lambda}{n})^n &= \sum_{i=0}^{n}\frac{n!}{i!(n-i)!}(-\frac{\lambda}{n})^i \\ &= \sum_{i=0}^{n}\frac{n!}{(n-i)!n^i}(-1)^i\frac{\lambda^i}{i!}\end{aligned}$$

Les termes $\frac{n!}{(n-i)!n^i}$ sont analogues au terme (3), ils convergent donc tous vers 1. Donc pour $n \to +\infty$ le terme (1) converge vers le développement en série entière de $e^{-\lambda}$ qui vaut :

$$e^{-\lambda} = \sum_{i=0}^{+\infty}\frac{(-\lambda)^i}{i!}$$

On a ainsi démontré que pour $n \to +\infty$,

$$\mathbb{P}[X = k] \to \frac{\lambda^k}{k!}e^{-\lambda}$$

Exercice 8. Le séquençage d'un génome ou d'un transcriptome peut être vu en première approximation comme un procédé stochastique dans lequel on tire successivement des fragments d'ADN dans un *pool* infini. Soit une région (ou un gène) d'intérêt et p la fraction des fragments du *pool* provenant de cette région.

1. Si on effectue un séquençage de profondeur n (*i.e.* on séquence n lectures), quelle est la distribution du nombre K de fragments provenant de la région d'intérêt ?

2. Expliquez pourquoi on peut approximer la distribution de K par une loi de Poisson dont vous préciserez le paramètre. Illustrez numériquement la qualité de l'approximation sous R.

3. Les programmes développés pour l'analyse des données de séquençage préfèrent systématiquement la loi de Poisson à la loi binomiale pour modéliser les comptages de lectures. Pouvez-vous expliquer cette préférence ?

1. On suppose les tirages successifs indépendants. Cela est notamment possible parce que l'on suppose le *pool* de fragments infini, donc les tirages effectués ne modifient pas sa composition, et le séquençage est assimilable à un tirage avec remise. Dans ce cas, chaque tirage est une variable de Bernoulli indépendante dont la probabilité de succès est p. On en déduit $K \sim \text{Binom}(n, p)$.

2. On sait que pour $n \to +\infty$, $p \to 0$ avec np constant, la loi $\text{Binom}(n, p)$ converge vers une loi de Poisson de paramètre np. Or pour une application typique de séquençage, la profondeur n est de l'ordre de 10^6-10^7, et la proportion p de lectures provenant d'une région d'intérêt comme un gène sera telle que np sera compris entre quelques unités et quelques milliers. Dans ce régime, l'approximation est déjà excellente, comme on peut le voir numériquement sous R :

```
> f <- function(n,p) { dbinom(n*p,n,p) / dpois(n*p,n*p) }
> f(10,0.1)
[1] 1.053118
> f(100,0.01)
[1] 1.005029
> f(1000,0.001)
[1] 1.0005
> f(10000,0.0001)
[1] 1.00005
```

2 Quelques exercices de plus

Exercice 9.

1. Développez les expressions $(x+y)^2$, $(x+y)^3$ et $(x+y)^4$.
2. On rappelle la formule du binôme de Newton :

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

Vérifiez que votre résultat de la question précédente est en accord avec la formule générale.

3. Démontrez par récurrence la formule du binôme de Newton.

1. facile !
2. facile !
3. Moins facile !
Pour $n = 0$: $(x+y)^0 = 1$ et $\sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} = \binom{0}{0} x^0 y^0 = 1$

Supposons la formule vraie pour n

$$
\begin{aligned}
(x+y)^{n+1} &= (x+y)(x+y)^n \\
&= (x+y)\sum_{k=0}^{n}\binom{n}{k}x^k y^{n-k} \quad \text{(hypothèse de récurrence)} \\
&= \sum_{k=0}^{n}\binom{n}{k}x^{k+1}y^{n-k} + \sum_{k=0}^{n}\binom{n}{k}x^k y^{n+1-k} \\
&= \sum_{k=1}^{n+1}\binom{n}{k-1}x^k y^{n+1-k} + \sum_{k=0}^{n}\binom{n}{k}x^k y^{n+1-k} \quad \text{(décalage d'indice dans la première somme)} \\
&= x^{n+1} + \sum_{k=1}^{n}\binom{n}{k-1}x^k y^{n+1-k} + y^{n+1} + \sum_{k=1}^{n}\binom{n}{k}x^k y^{n+1-k} \\
&= x^{n+1} + y^{n+1} + \sum_{k=1}^{n}\left(\binom{n}{k-1} + \binom{n}{k}\right)x^k y^{n+1-k} \\
&= x^{n+1} + y^{n+1} + \sum_{k=1}^{n}\binom{n+1}{k}x^k y^{n+1-k} \\
&= \sum_{k=0}^{n+1}\binom{n+1}{k}x^k y^{n+1-k}
\end{aligned}
$$

ce qui démontre l'égalité pour $n+1$. On se convaincra de l'égalité :

$$\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k}$$

en se rappelant que $\binom{n}{k}$ est le nombre de façons de choisir k objets parmi n. Ainsi, pour choisir k objets parmi $n+1$, soit on prend le premier, soit on ne le prend pas ; dans le premier cas, il reste $k-1$ objets à choisir parmi n, dans le deuxième cas, il faudra encore choisir k parmi n.

Exercice 10. Démontrez les propriété suivantes de la loi de Poisson. Soit $X \sim \mathcal{P}(\lambda)$:

1. $\mathbb{E}[X] = \lambda$
2. $\mathbb{V}[X] = \lambda$
3. Soit $Y \sim \mathcal{P}(\mu)$ indépendante de X. Alors $X + Y$ est une loi de Poisson de paramètre $\lambda + \mu$.

Indications :

1. se souvenir que $e^x = \sum_{i=0}^{+\infty} \frac{x^i}{i!}$
2. utiliser la relation $\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
3. (re)voir la notion de produit de convolution et la formule du binôme de Newton.

$$\begin{aligned}
\mathbb{E}[X] = \sum_{i=0}^{+\infty} i\mathbb{P}[X=i] &= \sum_{i=0}^{+\infty} i\frac{\lambda^i}{i!}e^{-\lambda} &&= e^{-\lambda}\sum_{i=1}^{+\infty} i\frac{\lambda^i}{i!} \\
&= e^{-\lambda}\sum_{i=1}^{+\infty}\frac{\lambda^i}{(i-1)!} = e^{-\lambda}\lambda\sum_{i=1}^{+\infty}\frac{\lambda^{(i-1)}}{(i-1)!} &&= e^{-\lambda}\lambda\sum_{i=0}^{+\infty}\frac{\lambda^i}{i!} \\
&= e^{-\lambda}\lambda e^{\lambda} &&= \lambda
\end{aligned}$$

$$\begin{aligned}
\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 &= \sum_{i=0}^{+\infty} i^2\mathbb{P}[X=i] - \lambda^2 &&= \sum_{i=1}^{+\infty} i\frac{\lambda^i}{(i-1)!}e^{-\lambda} - \lambda^2 \\
&= \lambda\sum_{i=1}^{+\infty} i\frac{\lambda^{(i-1)}}{(i-1)!}e^{-\lambda} - \lambda^2 = \lambda\sum_{i=0}^{+\infty}(i+1)\frac{\lambda^i}{i!}e^{-\lambda} - \lambda^2 = \lambda\mathbb{E}[X+1] - \lambda^2 \\
&= \lambda(\mathbb{E}[X]+1) - \lambda^2 &&= \lambda
\end{aligned}$$

On pose $Z = X + Y$.

$$\begin{aligned}
\mathbb{P}[Z=k] &= \sum_{i=0}^{k}\mathbb{P}[X=i]\mathbb{P}[Y=k-i] \text{ (produit de convolution de deux variables aléatoires)} \\
&= \sum_{i=0}^{k}\frac{\lambda^i}{i!}\frac{\mu^{(k-i)}}{(k-i)!}e^{-\lambda-\mu} \\
&= \frac{e^{-\lambda-\mu}}{k!}\sum_{i=0}^{k}\frac{k!}{i!(k-i)!}\lambda^i\mu^{(k-i)} \\
&= \frac{e^{-\lambda-\mu}}{k!}(\lambda+\mu)^k
\end{aligned}$$

qui correspond bien à la distribution de la loi de Poisson de paramètre $\lambda + \mu$.

3 Application : modèle de survie en écotoxicologie

Les productions manufacturières incorporent de très nombreuses substances avec des motivations allant de la réduction des coûts à l'amélioration des produits en passant par l'attrait marketing, que ce soit pour l'alimentation, les textiles, l'électronique, *etc*. Ces substances peuvent présenter des risques non seulement pour les utilisateur·ice·s directs mais également pour les espèces des milieux naturels, lorsqu'elles y sont dispersées par des accidents, une mauvaise gestion des déchets ou les processus de production eux-mêmes. Pour anticiper et prévenir les conséquences des pollutions chimiques sur la biodiversité, les substances chimiques utilisées dans l'industrie sont progressivement testées pour évaluer quantitativement leurs effets sur différentes espèces animales et végétales. Nous nous intéressons ici à l'un des types de tests réalisés, le test de survie à terme fixé.

Essentiellement, il s'agit d'exposer un groupe d'individus à une dose contrôlée et constante d'une substance (contaminant) pendant un temps déterminé, et à compter le nombre d'individus survivants à la fin de la période considérée. Ce test est répété à différentes concentrations, et on souhaite en déduire une concentration seuil au-delà de laquelle le risque pour la survie des individus n'est plus négligeable.

1. La probabilité de survie d'un individu en fonction de la concentration présente de contaminant est très souvent modélisée à l'aide d'une fonction *logistique*, c'est-à-dire de la forme :

$$f(x) = \frac{1}{1 + e^{b(x-\theta)}}$$

Représentez graphiquement avec R une telle fonction, et expliquez ce que ce choix implique sur la réponse des individus à l'exposition au contaminant. Que représentent θ et b ?

2. Proposez un modèle probabiliste pour un test de survie comprenant r répétitions réalisées à différentes concentrations et avec différents nombres d'individus en début d'expérience.

3. Programmez en R un simulateur de votre modèle, c'est-à-dire une fonction prenant en entrée un tableau de concentrations testées, une valeur de θ et de b et générant des données expérimentales.

4. Comment représentez graphiquement un jeu de données de survie comme fourni par votre simulateur ? Proposez une implémentation, puis complétez-la avec la possibilité de tracer la "vraie" courbe de probabilité de survie en fonction de la concentration de contaminant.

5. Écrivez la fonction de vraisemblance de votre modèle, puis la log-vraisemblance.

6. Implémentez en R le calcul de la log-vraisemblance et proposez un petit test pour vous convaincre de la justesse de votre implémentation.

7. Il n'existe pas dans ce cas de formule analytique pour calculer le maximum de la log-vraisemblance, mais nous pouvons l'approcher par optimisation numérique. Montrez comment réaliser cela à l'aide la fonction `optim` de R.

8. Représentez sur un même graphique un jeu de données de survie généré par votre simulateur, la courbe de survie pour les paramètres ayant servis à la simulation, et la courbe estimée en maximum de vraisemblance.

9. Comment pourrait-on modifier le modèle pour tenir compte de la mortalité "naturelle" des individus durant l'essai ?

1. Le code R pour représenter la fonction

```
sigmoid <- function(theta, b, x) {
    1 / (1 + exp(b * (x - theta)))
}

plot_sigmoid <- function(theta, b, xmin, xmax, col = "black", add=FALSE) {
  f <- function(x) {sigmoid(theta, b,x)}
  x <- seq(xmin, xmax,length.out=100)
  y <- sapply(x, f)
  if(add) { lines(x, y, col=col) }
  else { plot(x, y,type='l', col=col) }
```

```
}
plot_sigmoid(-3,1,-10,3)
plot_sigmoid(-6,1,-10,3,add=T,col="red")
plot_sigmoid(-3,3,-10,3,add =T, col="blue")
```

La forme de la courbe implique qu'à dose faible la présence du contaminant n'a aucune incidence sur la mortalité des individus. Lorsque la dose augmente, on aura un effet seuil, c'est-à-dire une dose au-dessus de laquelle l'effet augmentera brutalement, pour arriver à une survie nulle. On observe en effet que $\lim_{x\to-\infty} f(x) = 1$ et $\lim_{x\to+\infty} f(x) = 0$. Ici l'énoncé aurait dû préciser que l'on travaille en général en échelle logarithmique (i.e., x est le logarithme de la concentration), ce qui fait que l'on peut avoir des valeurs négatives comme positives. Par ailleurs, $f(\theta) = \frac{1}{2}$, θ correspond à la dose à laquelle la survie est de 50%, au milieu de la zone seuil. Lorsqu'on diminue ou on augmente θ la courbe est décalée sur la gauche ou la droite respectivement. Le paramètre b contrôle la pente de la courbe dans la zone seuil, c'est-à-dire la vitesse avec laquelle on passe d'une probabilité de survie 1 à une probabilité nulle.

2. Ici l'énoncé manquait de clarté. Pour fixer les idées, disons qu'on se propose de mettre en place r aquariums dans lesquels on place un nombre éventuellement à chaque fois différents d'individus exposés à une certaine dose du contaminant, éventuellement variable à chaque fois. On pose le modèle suivant :

θ, b : paramètre de la fonction de survie pour l'espèce et le contaminant considérés
r : nombre d'aquariums
n_i : nombre d'individus initialement présents dans l'aquarium i
c_i : concentration de contaminant dans l'aquarium i
k_i : nombre de survivants observés dans l'aquarium i à l'issue de l'expérience
$k_i \sim \text{Binom}(n_i, f(c_i))$

On choisit ici une distribution binomiale pour k_i, qui suppose que la survie d'un individu dans un aquarium n'a pas d'influence sur celle de ses colocataires, et que toutes les individus présentent la même sensibilité au contaminant.

3.
```
simulateur <- function(n_init, tested_concentrations, theta, b) {
    rbinom(length(n_init),
           n_init,
           sapply(tested_concentrations, function(c) sigmoid(theta, b, c)))
}
simulateur(c(4,8,3), c(-4,-1,2), -3, 1)
```

4. Comme chaque expérience part possiblement d'un nombre d'individus différent, il faut représenter la proportion observée plutôt que le nombre de survivants.

```
plot_obs  <- function(n_init, tested_concentrations, obs, theta, b) {
    plot(tested_concentrations, obs / n_init, pch=19)
    plot_sigmoid(theta, b,
```

```
                min(tested_concentrations), max(tested_concentrations),
                col="gray", add=T)
}

demo_plot_obs <- function(n_init, tested_concentrations, theta, b) {
    obs  <- simulateur(n_init,tested_concentrations,theta,b)
    plot_obs(n_init,tested_concentrations,obs,theta,b)
}

demo_plot_obs(c(4,8,4,5,3), c(-4,-3,-1,0,2), -3, 1)
```

5. Vraisemblance

$$\begin{aligned} L(\theta, b \mid k_1, \ldots, k_n) &= \mathbb{P}[k_1, \ldots, k_n \mid theta, b] \\ &= \prod_{i=1}^{n} \mathbb{P}[k_i \mid theta, b] \\ &= \prod_{i=1}^{n} \binom{k_i}{n_i} f(c_i)^{k_i} (1 - f(c_i))^{n_i - k_i} \end{aligned}$$

Log-vraisemblance

$$\log L(\theta, b \mid k_1, \ldots, k_n) = \sum_{i=1}^{n} \left(\log \binom{k_i}{n_i} + k_i \log f(c_i) + (n_i - k_i) \log(1 - f(c_i)) \right)$$

On remarque que les termes log binomiaux ne dépendent pas des paramètres, on pourra donc les ignorer pour l'optimisation.

6.
```
loglikelihood <- function(n_init, tested_concentrations, theta, b, obs) {
    sum(
        sapply(1:length(obs), function(i) {
            p_i <- sigmoid(theta, b, tested_concentrations[i])
            log(dbinom(obs[i], n_init[i], p_i))
        }))
}

> obs <- simulateur(c(4,8,4,5,3,10,12), c(-4,-3,-1,0,2,-2,-5), -3, 1)
> plot_obs(c(4,8,4,5,3,10,12), c(-4,-3,-1,0,2,-2,-5), obs, -3, 1)
> loglikelihood(c(4,8,4,5,3,10,12), c(-4,-3,-1,0,2,-2,-5), -3, 1, obs)
[1] -7.922754
> loglikelihood(c(4,8,4,5,3,10,12), c(-4,-3,-1,0,2,-2,-5), -2, 2, obs)
[1] -24.34665
> loglikelihood(c(4,8,4,5,3,10,12), c(-4,-3,-1,0,2,-2,-5), -5, 0.1, obs)
[1] -19.354
```

Le test montre que pour un jeu de données simulées, les paramètres ayant servi à la simulation sont plus vraisemblables que les deux autres que l'on a testé : ça va dans le bon sens.

7.
```
ml_inference <- function(n_init, tested_concentrations,obs) {
    optim(c(0, 1),
          function(x) -loglikelihood(n_init,
                                     tested_concentrations,
                                     x[1],x[2],obs))$par
}
> ml_inference(c(4,8,4,5,3,10,12), c(-4,-3,-1,0,2,-2,-5), obs)
[1] -3.659787  1.786261
```

8.
```
demo_inference <- function(n_init, tested_concentrations, theta, b) {
    obs  <- simulateur(n_init,tested_concentrations,theta,b)
    plot_obs(n_init,tested_concentrations,obs,theta,b)
    p_hat <- ml_inference(n_init,tested_concentrations,obs)
    plot_sigmoid(p_hat[1], p_hat[2],
                 min(tested_concentrations), max(tested_concentrations),
                 col="red", add=T)
}

demo_inference(c(4,8,4,5,3,10,12), c(-4,-3,-1,0,2,-2,-5),-3,1)
```

9. Il suffirait d'introduire un nouveau paramètre a dans la fonction de survie

$$f(x) = \frac{a}{1 + e^{b(x-\theta)}}$$

et de l'estimer comme les autres paramètres.

4 Le mot de la fin

Qu'est-ce que ça veut dire concrètement, modéliser ? Lorsque l'on réalise une expérience, on obtient en bout de course des observations. Dans certains cas, il est aisé d'en déduire la conséquence qui motivait l'expérience (p. ex., déterminer si une protéine est exprimée à l'aide d'un *Western blot*). Dans d'autres, l'information apportée par ces observations est soit indirecte, soit quantitative, et il faut alors recourir à des outils formels pour tirer les conséquences de ce que l'on observe.

On procède en trois étapes :

1. lister les variables utiles pour décrire l'expérience, en introduisant une notation pour chacune d'elles
2. expliciter comment ces différentes variables sont liées les unes aux autres
3. dériver, à travers ces liens, les conséquences induites par les observations.

La forme prise par le modèle dépend du cadre mathématique dans lequel on se place : modèle probabiliste, équations différentielles, systèmes de contraintes ... Dans les modèles que nous étudions ce semestre, les variables sont décrites soit comme des fonctions déterministes des

autres, soit comme des variables aléatoires dont la distribution dépend des autres variables. Les observations seront utilisées pour soit estimer les paramètres du modèle par maximum de vraisemblance, soit pour dériver une distribution sur les paramètres en utilisant le théorème de Bayes.

La notion de paramètre Parmi les variables que l'on est amené à définir dans un modèle, certaines sont de nature à changer de valeur chaque fois que l'on refait l'expérience, d'autres au contraire sont une propriété intrinsèque du système que l'on étudie et ne changent pas lorsqu'on refait l'expérience. Ces dernières sont appelées paramètres du modèle, et ce sont en général elles que l'on cherche à estimer à partir des observations. Par exemple, si l'on étudie la prévalence d'une maladie dans une population, le nombre d'individus malades dans l'échantillon constitué pourra varier quand on réplique l'expérience, mais d'une fois sur l'autre la prévalence sera supposée constante et constituera le paramètre à estimer.

Doit-on spécifier dans le modèle une distribution de probabilité pour les paramètres ? Deux approches sont possibles. La première, dite approche fréquentiste, considère qu'un modèle probabiliste est une recette expliquant les fréquences auxquelles on observerait les différents résultats possibles d'une expérience si on la répétait à l'identique un grand nombre de fois. Les paramètres sont simplement des inconnues de cette recette, et on les trouvera en cherchant la valeur qui rendra nos observations les plus probables : c'est le principe du maximum de vraisemblance. Dans ce cadre, le maximum de vraisemblance est donc le levier qui transfère l'information des données observées à nos paramètres et on ne spécifie pas de distribution de probabilité pour nos paramètres.

La deuxième approche, dite bayésienne, utilise la notion de probabilité pour représenter non pas la fréquence d'un évènement mais notre incertitude sur la valeur d'une variable. Dans ce cadre, on spécifie une distribution de probabilité pour toutes les variables, paramètres y compris. Pour les paramètres, cette distribution est dite *a priori*, c'est-à-dire qu'elle représente notre incertitude sur les paramètres *avant* d'avoir réalisé notre expérience. Pour exploiter les résultats de l'expérience, on calcule à l'aide du théorème de Bayes la distribution de probabilité sur les paramètres *sachant les données*, dite distribution *a posteriori*, qui représente notre incertitude sur la valeur des paramètres *après* avoir intégré les résultats de l'expérience. Ici, c'est le théorème de Bayes qui permet le transfert de l'information contenue dans les observations vers les paramètres.