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

Philippe Veber

22 janvier 2020

1 Loi de Bernoulli

On appelle **épreuve de Bernoulli** un évènement qui peut ou non se réaliser (on parle de succès s'il se réalise). Pour représenter la survenue de l'évènement (ou le succès de l'épreuve), on introduit une v.a. X qui prend la valeur 1 avec une probabilité p et 0 avec probabilité $1 - p$. La loi suivie par cette v.a. est appelée **loi de Bernoulli** de paramètre p et est notée $\mathrm{Bern}(p)$.

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.

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

2 Loi binomiale

Soit une collection de v.a. $X_1, \ldots, X_n$ indépendantes et suivant une loi de Bernoulli de paramètre p. Alors la loi suivie par

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

est appelée **loi binomiale**. Comme chaque variable X_i représente le succès d'un test, la variable K compte le nombre de tests positifs dans la collection. Il s'agit donc d'une v.a. à valeurs entières, comprises entre 0 et n, et on la note $\mathrm{Binom}(n, p)$. La distribution de probabilité est donnée par la formule :

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

Le coefficient binomial $\binom{n}{k}$ représente le nombre de façons d'obtenir k succès parmi n tests ; et chacune d'elle a la probabilité $p^k(1-p)^{(n-k)}$.

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 ?

3 Variable catégorielle

Une variable catégorielle est une v.a. qui ne peut prendre qu'un nombre fini de valeurs. Une variable suivant une loi de Bernoulli est une variable catégorielle, puisqu'elle ne peut prendre que deux valeurs. Autres exemples, l'acide aminé se trouvant à une position donnée d'une protéine (une valeur parmi 20), une catégorie d'âge, de sexe … La loi de probabilité d'une variable catégorielle à n valeurs peut être entièrement décrite par n paramètres $p_1, \ldots, p_n$ tels que $\forall i \in \{1, \ldots, n\}\ p_i \in [0;1]$ et $\sum_{i=1}^{n} p_i = 1$. On remarque que puisque la somme de ces paramètres doit être égale à 1, on peut se contenter d'en donner $n-1$ (pour la loi de Bernoulli, on ne précise ainsi que la probabilité p de la valeur 1).

Une v.a. X catégorielle sur un ensemble fini $\mathcal{A} = \{a_1, \ldots, a_n\}$ sera notée $X \sim \mathrm{Cat}(\mathcal{A}, p)$, où p désigne le vecteur des probabilités $p_i = \mathbb{P}[X = a_i]$. Si $\mathcal{A} = \{1, \ldots, n\}$, ou si l'identité de $\mathcal{A}$ est évidente d'après le contexte, on s'autorisera à dire $X \sim \mathrm{Cat}(p)$.

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 :

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

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.

4 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 ?

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 ?

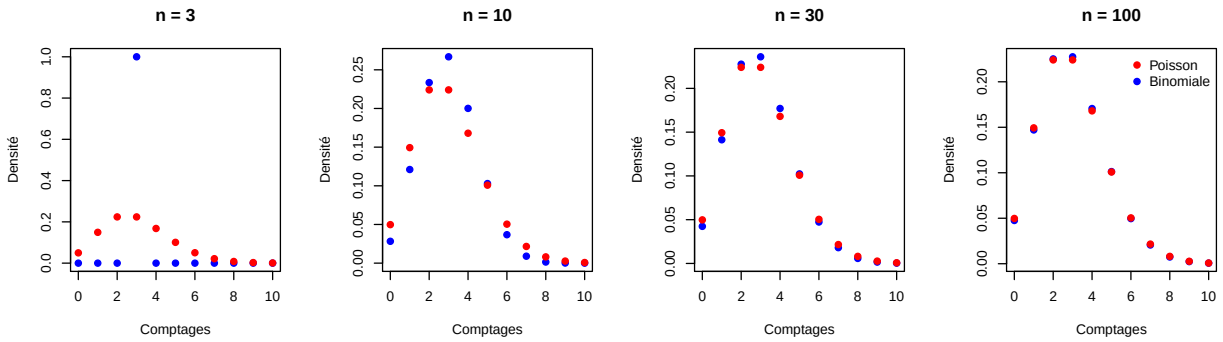

FIGURE 1 – Loi binomiale et loi de Poisson. Les graphes représentent en bleu la même distribution, la loi de Poisson de paramètre 3 ; en rouge, la distribution de lois binomiales de paramètre n et $\frac{3}{n}$.

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.

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 ?

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

5 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 ?