TP2 : loi de l'EMV

marie-anne.poursat@universite-paris-saclay.fr

30 janvier 2024

Objectifs du TP :

- Estimer par MV une densité paramétrique et visualiser l'ajustement par un qqplot;
- estimer l'écart-type de l'EMV et calculer un intervalle de confiance du paramètre.

1 Ajustement de données de comptages

Le tableau de données suivant reprend les observations d'une expérience de physique menée par Berkson 1966, et reprise par John Rice (Mathematical Statistics and Data Analysis, 1995, p.240). Pendant 1207 intervalles de temps consécutifs de 10 secondes, un expérimentateur a observé le nombre d'émissions de particules alpha. Les comptages Y représentent le nombre d'intervalles de temps au cours desquels ont été observés n émissions. Ainsi au cours de 18 intervalles de temps parmi les 1207, il y a eu 2 émissions de particules, au cours de 28 intervalles de temps 3 émissions etc... On dit que les données sont *agrégées* ou *répétées*.

n	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Y	18	28	56	105	126	146	164	161	123	101	74	53	23	15	9	5

Modélisation des données : soit X_i le nombre d'émissions de particules observé pendant l'intervalle de temps i; on considère que les $X_i,\ i = 1, \ldots, 1207$ sont des variables indépendantes suivant une loi de Poisson d'espérance λ; les comptages observés $Y_n,\ n = 2, \ldots, 17$ représentent les comptages agrégés selon le nombre d'émissions de particules.

1.1 Calcul de l'EMV

1. Lire le jeu de données `counts.txt` et créer le vecteur `y` des comptages agrégés. Vérifier qu'il est de longueur 16.
   Représenter un diagramme en bâtons des comptages.
2. Estimer par maximum de vraisemblance le paramètre λ, c'est à dire le nombre moyen d'émissions par intervalle de temps.
3. Superposer sur le graphe précédent l'ajustement par maximum de vraisemblance de la loi de Poisson à ces données.
   Renseigner les axes et ajouter un titre.

1.2 qqplot

Lorsque l'on compare deux distributions (celles de deux échantillons, ou bien la distribution d'un échantillon avec une distribution théorique ou estimée), les graphiques des histogrammes ne sont pas (visuellement) informatifs.

Le graphe quantile/quantile ou Q-Q plot est un outil pour vérifier visuellement l'adéquation d'un jeu de données à une loi. Il représente les **quantiles théoriques** du modèle, associés à la loi estimée, en fonction des **quantiles empiriques**, observés sur l'échantillon. Si la loi théorique est pertinente, les points s'alignent à peu près, pas exactement!) suivant la première diagonale.

En pratique, il y a plusieurs façons de calculer les quantiles empiriques de l'échantillon observé. L'implémentation peut différer légèrement selon les logiciels, qui prévoient d'ailleurs plusieurs types de calcul (argument `type` de la fonction `quantile` de R).

```
plot(qpois(ppoints(n),...),quantile(...,ppoints(n)),col="blue",
     xlab="quantile théorique", ylab="quantile empirique") # à compléter
abline(0,1)
```

4. Comprendre les fonctions R `ppoints` et `quantile`.
5. Vérifier l'adéquation du modèle aux données avec un graphe Quantile-Quantile.

1.3 IC de l'EMV

6. Tracer la log-vraisemblance de l'échantillon.
7. Calculer l'écart-type estimé de l'EMV puis un intervalle de confiance de niveau 95 % de λ.
   Repérer l'IC sur le tracé de la log-vraisemblance (par exemple avec `segments`).
8. Retrouver les résultats précédents (valeur et écart-type estimé de l'EMV) en utilisant l'algorithme d'optimisation `optim`.

Dans les modèles des TP suivants, on ne pourra pas comme ici calculer "à la main" l'EMV et son écart-type, on utilisera donc des algorithmes d'optimisation pour ajuster le modèle aux données.

2 Pour approfondir: étude par simulation de la loi de l'EMV

1. Générez `S`=200 échantillons (`replicate()`) $Y^s, s = 1, \ldots, 200$ de taille $n = 20$ et de loi de Poisson d'espérance $\lambda = 3$.
   Calculez les valeurs de l'EMV pour chacun de ces échantillons; ce vecteur est appelé *loi d'échantillonnage* de l'EMV. Représentez graphiquement par un histogramme cette loi d'échantillonage.
2. Superposez à l'histogramme précédent la loi gaussienne de moyenne λ et de variance $1/I(\lambda)$ où $I(\lambda)$ est l'information de Fisher du modèle associée à un échantillon de taille n. Ici nous connaissons la loi générative (étude de simulation) et nous pouvons calculer l'information de Fisher théorique.
3. Tracez la fonction de répartition empirique de la loi d'échantillonage de l'EMV et comparer avec la loi asymptotique gaussienne.
4. Commentez la qualité de l'ajustement de la loi gaussienne à la loi de l'EMV par un qqplot. Refaites la simulation en faisant varier n.