Étude de la distribution du nombre de transcrits par gouttelettes

Julien Heurtin - **Louis Allain** - **Léonard Gousset**

Groupe 37 - Mai 2023

Tuteur : Emmanuel Curris

ENSAI — École nationale de la statistique et de l'analyse de l'information

UNIVERSITÉ PARIS DESCARTES

Introduction

Objectif

- Etudier la régénération des cellules souches après une greffe dans le cas d'une leucémie.

Comment faire ?

- Etudier le transcriptome (ensemble des ARN messagers) de chaque cellule, grâce à la "single cell RNA-seq". → Diviser les cellules en suspension en gouttelettes individuelles
- Analyser le contenu de ces gouttelettes avec un modèle de mélange.

Nos données

- Le nombre de transcrits dans chaque gouttelette. Dont le support est à valeurs dans $\mathbf{N}^*$.

Méthodes

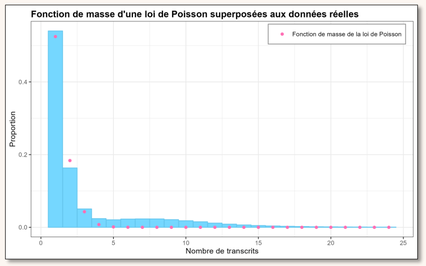
Un modèle de mélange peu usuel

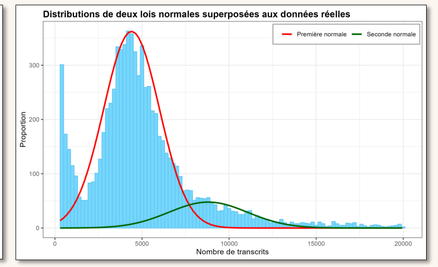
- Modèle de mélange à composantes discrètes et continues
  - Une loi de Poisson modélisant les gouttelettes contenant des débris cellulaires.
  - Une première loi normale modélisant les gouttelettes à une seule cellule.
  - Une seconde loi normale modélisant les gouttelettes à 2 cellules.
- La correction de continuité permet d'approcher nos données discrètes par des lois continues.
- Pour trouver les paramètres optimaux (λ, μ et σ), nous avons mis en place un **algorithme EM** (Espérance-Maximisation).

Résultats attendus

Une approche exploratoire prometteuse

Selon les hypothèses du sujet, on pourrait représenter nos données uniquement par une loi de Poisson et deux lois normales. La superposition de la fonction de masse (resp. Des densités) pour la loi de Poisson (resp. Les lois normales) sur nos données pour voir les résultats.





- La loi de Poisson représente bien le début de nos données. On observe tout de même une bosse entre 5 et 15 transcrits qui n'est pas modélisée par la loi.
- Les deux lois normales sont-elles, très encourageantes. Les gouttelettes contenant une seule cellule sont bien modélisées, tout comme celles à 2 cellules.

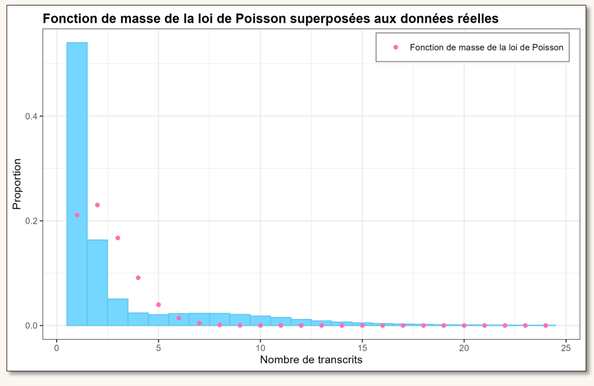
Résultats obtenus

Un modèle à améliorer

Nous obtenons les résultats suivants :

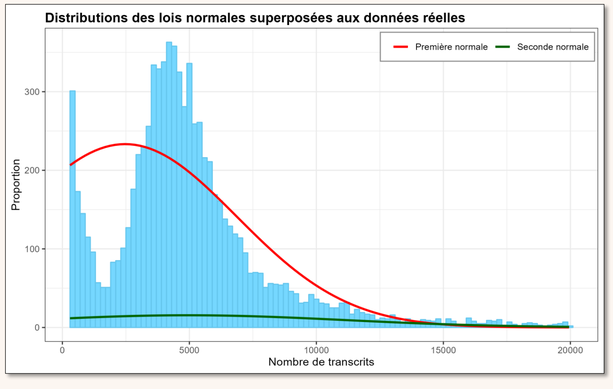
	Groupe 1	Groupe 2	Groupe 3
Minimum	1	12	15 558
Moyenne	2,4 (+2,5)	1 550 (+2 714)	3313 (+12039)
Maximum	11	15696	72639
Population	301921	24963	511
Proportion	92,2%	7,6%	0,2%

En superposant la fonction de masse de la loi de Poisson (simulée avec le λ trouvé par notre modèle) sur nos données réelles, on retrouve le graphique suivant :



La loi de Poisson modélise incorrectement les données, sous-estimant les gouttelettes à un seul transcrit, surestimant les gouttelettes de 2 à 5 transcrits et mal modélisant les gouttelettes formant une **bosse** entre 6 et 15 transcrits.

Les conséquences sont directes pour le paramètre μ qui est donc trop faible du fait des gouttelettes comprises entre 6 et 15 transcrits non représentés par notre loi de Poisson et que la loi normale doit donc inclure dans le groupe 2.



La variance est également très élevée et peu représentative de nos données comme on peut le voir sur le graphique. Le groupe 3 dépendant des paramètres du groupe 2 est donc lui aussi **mal représenté** par conséquence.

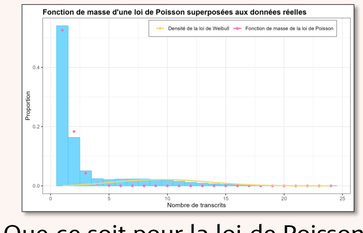
Conclusion

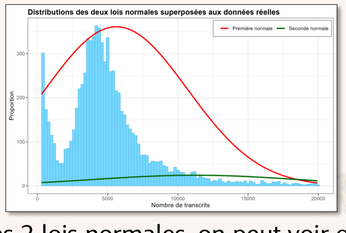
Un modèle incomplet

Bien que les résultats de l'algorithme EM soient encourageants, notre modèle n'arrive pas à représenter correctement les données. On peut expliquer cela par les hypothèses des lois. En effet, dans la partie résultats attendus, on retrouve une partie des données qui n'est pas modélisée. Un travail de recherche est encore nécessaire et une exploration plus profonde du modèle serait intéressante.

Pour aller plus loin

Nous avons tenté de modéliser la "bosse" énoncée dans les parties précédente par une loi de Weibull dont voici les résultats :





Que ce soit pour la loi de Poisson ou les 2 lois normales, on peut voir que les résultats sont plus concluants et l'idée d'explorer le modèle est toujours plus forte.