Promenade mathématique : Statistique, santé et biologie

Agathe Guilloux
Professeure à l'Université d'Évry - Paris Saclay

Introduction : statistique, santé et biologie

La statistique

La statistique est la **science des données** :

à partir de

- **données** et
- d'un **modèle probabiliste**
- on veut **décrire** et **vérifier des hypothèses**.

Quelques exemples en santé

- essais cliniques
  - traitement versus placebo
  - les malades ont-ils guéris ?
- marqueurs thérapeutiques
  - à partir de données génétiques (génomiques)
  - peut-on prévoir à quel traitement le patient répondra le mieux ?
- épidémiologie
  - à partir de tweets
  - peut-on repérer le début d'une épidémie ?

Un exemple en biologie

Cellules du sang

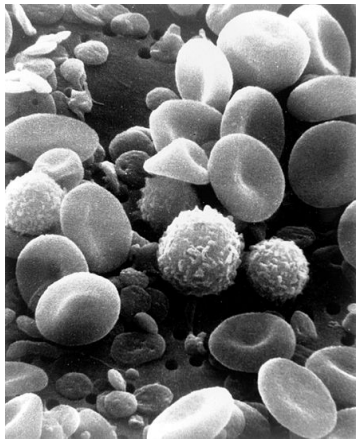

FIG. : Globules rouges, des lymphocytes, un monocyte, un neutrophile, plusieurs plaquettes... [Wik17b]

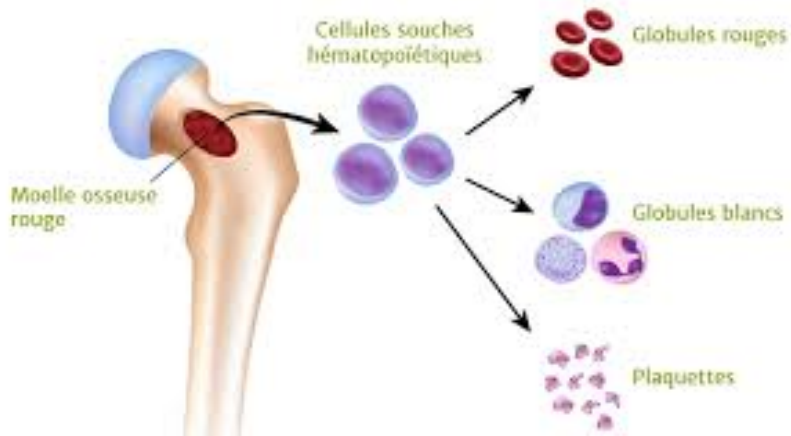

FIG.: L'hématopoïèse

Cellule souche et hématopoïèse

- Une cellule souche est une **cellule indifférenciée** capable, à la fois, de **générer des cellules spécialisées** [...] et de **se maintenir** dans l'organisme [...]. (cf. [Wik17a])
- L'**hématopoïèse** est le processus par lequel toutes les **cellules sanguines** sont produites.

On sait depuis le début des années 60 qu'il existe une **cellule souche hématopoïétique** [BMT63]. Elle capable de régénérer l'ensemble des cellules du sang .

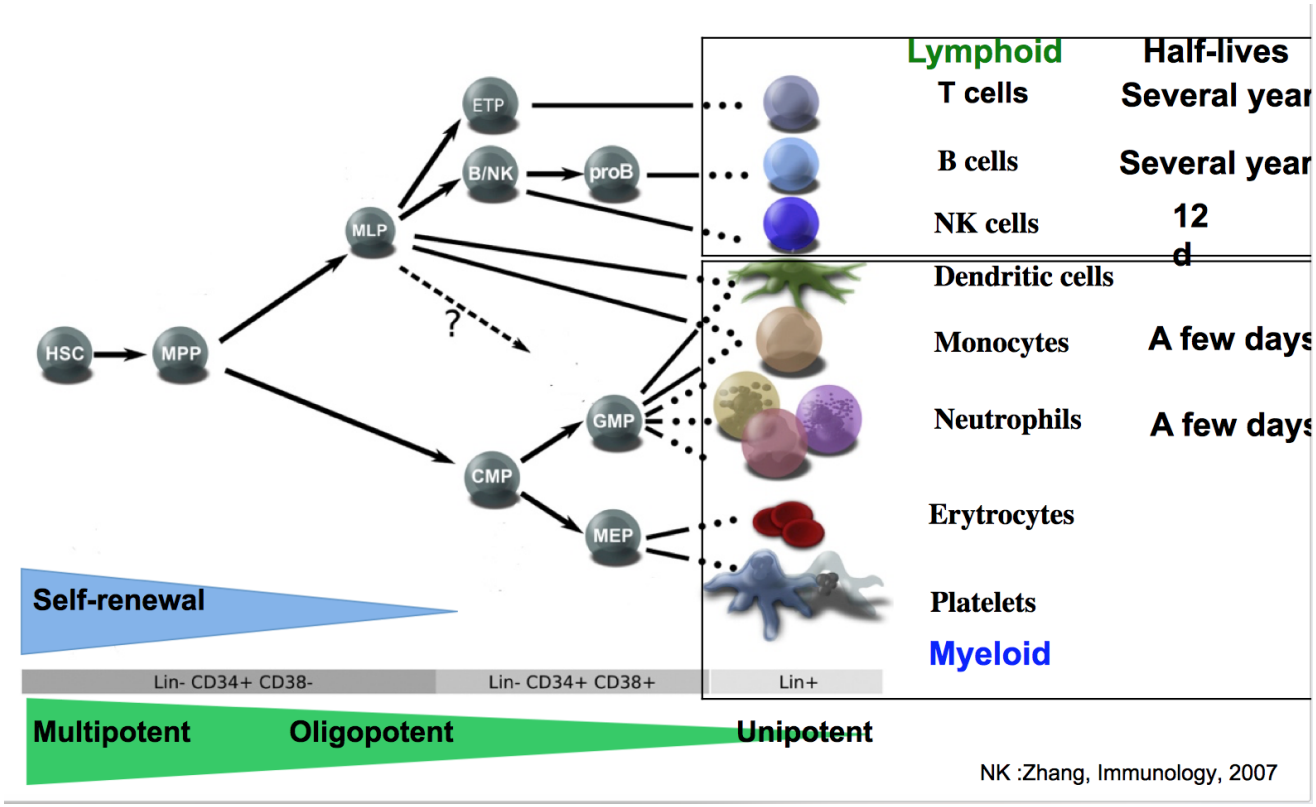

FIG.: L'hématopoïèse

Une hypothèse

Le modèle historique

Il existe un **unique type** de cellule souche hématopoïétique.

Conséquence : toutes les cellules souches doivent produire les différentes cellules sanguines en même proportion.

Ici pour les cellules qui nous intéressent (du système immunitaire)

B (type 1)	M (type 2)	N (type 3)	T (type 4)	NK (type 5)
0.266	0.114	0.27	0.076	0.274

Dé pipé et données pour $n = 10$ dans le sang

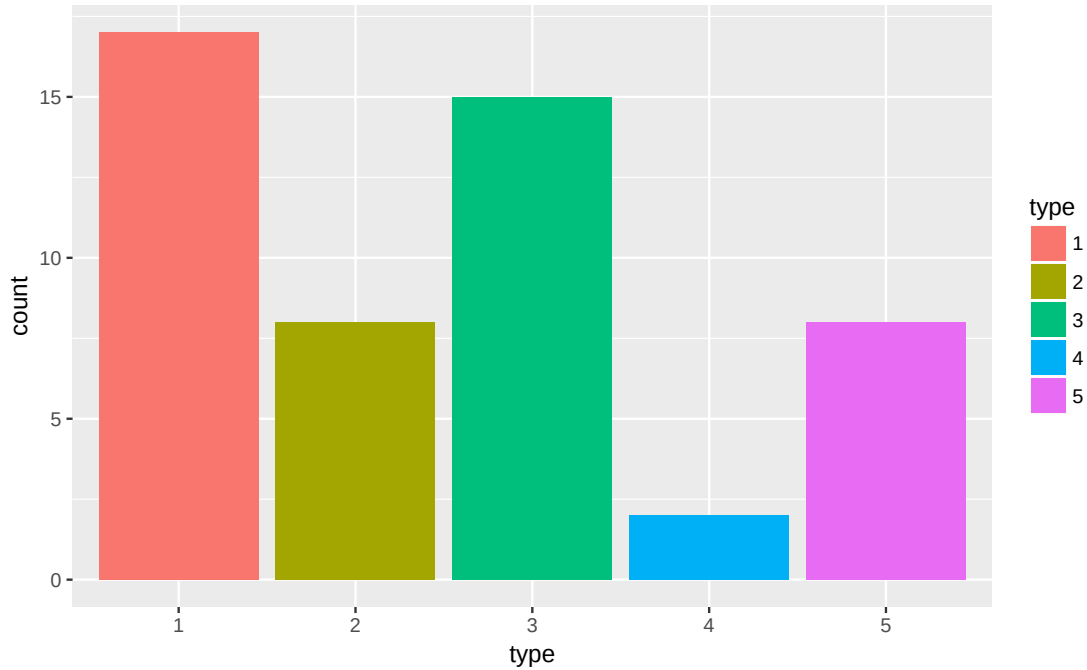

FIG.: Observation pour 5 cellules souches et 10 cellules différenciées

B (type 1)	M (type 2)	N (type 3)	T (type 4)	NK (type 5)
0.266	0.114	0.27	0.076	0.274

Dé pipé et données pour $n = 10$ pour chaque CS

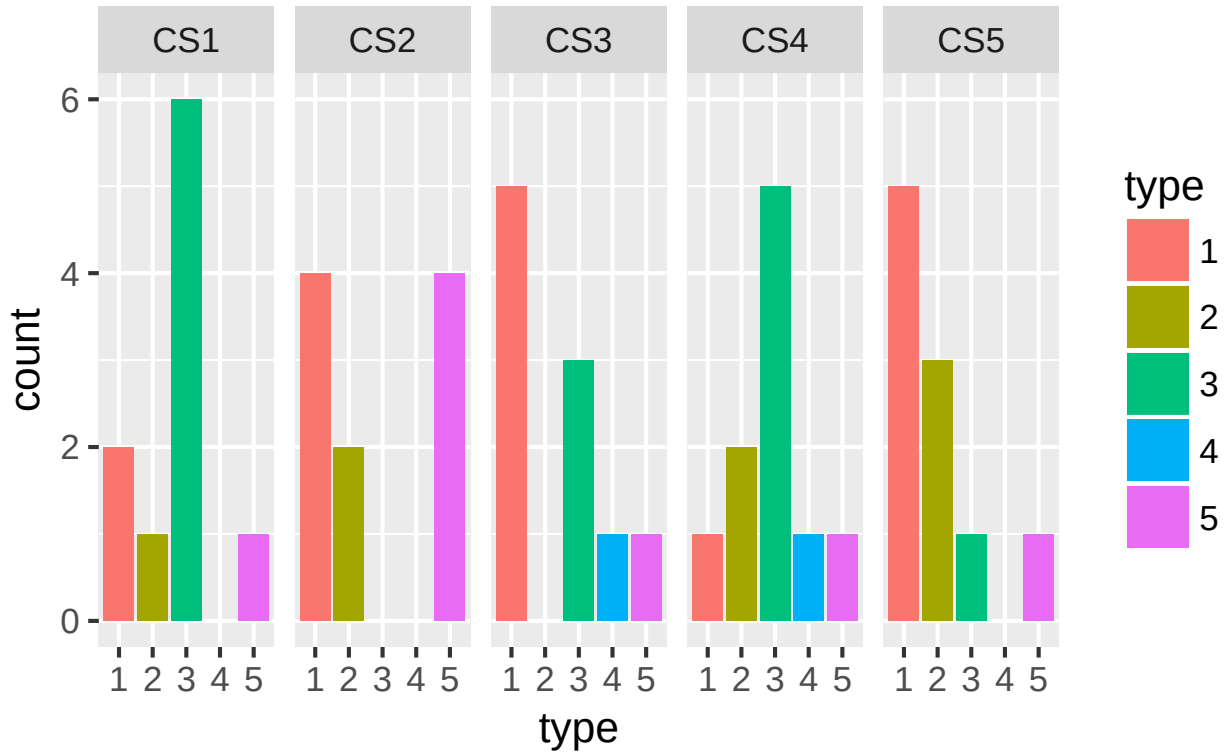

FIG.: Observation pour 5 cellules souches et 10 cellules différenciées

Dé pipé et données pour $n = 1000$ dans le sang

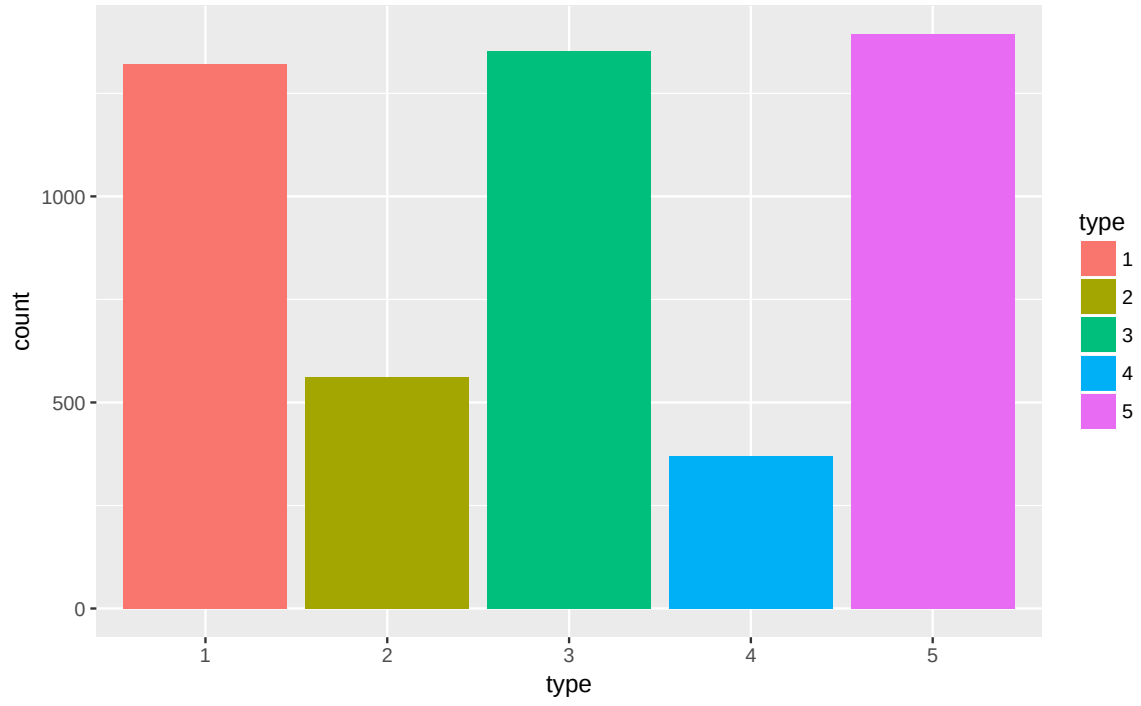

FIG.: Observation pour 5 cellules souches et 1000 cellules différenciées

B (type 1)	M (type 2)	N (type 3)	T (type 4)	NK (type 5)
0.266	0.114	0.27	0.076	0.274

Dé pipé et données pour $n = 1000$ par cellule souche

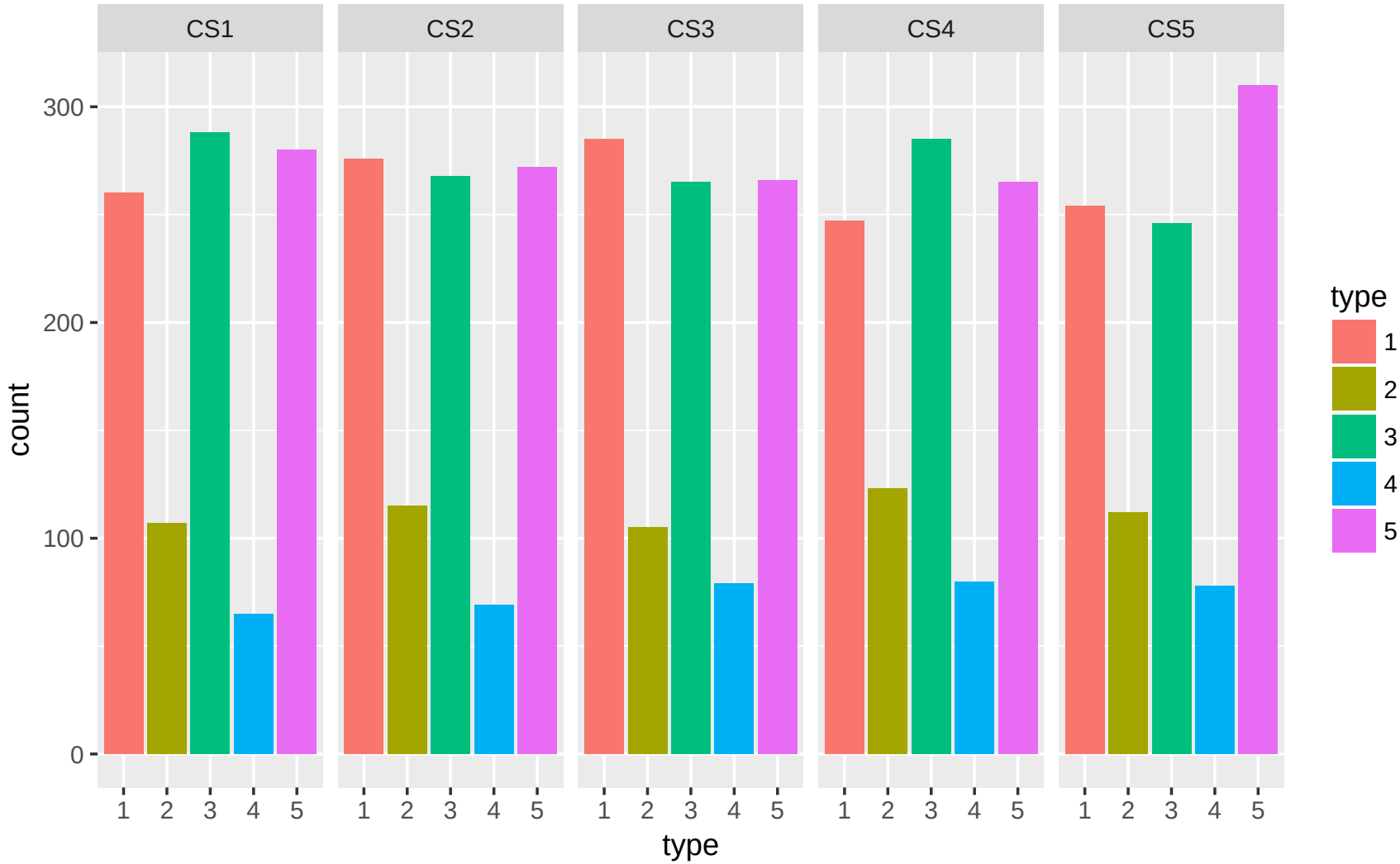

FIG.: Observation pour 5 cellules souches et 1000 cellules différenciées

Et si c'était faux ? ? ?

Des biologistes ont avancé l'hypothèse de l'existence de **plusieurs types de cellules souches en terme de potentialité** [DKB+07].

Un exemple

Chaque cellule souche produit

- avec **probabilité** $2/5$

B (type 1)	M (type 2)	N (type 3)	T (type 4)	NK (type 5)
0.41	0.17	0.41	0.00	0.00

- avec **probabilité** $3/5$

B (type 1)	M (type 2)	N (type 3)	T (type 4)	NK (type 5)
0.00	0.00	0.00	0.22	0.78

Données : $n = 1000$

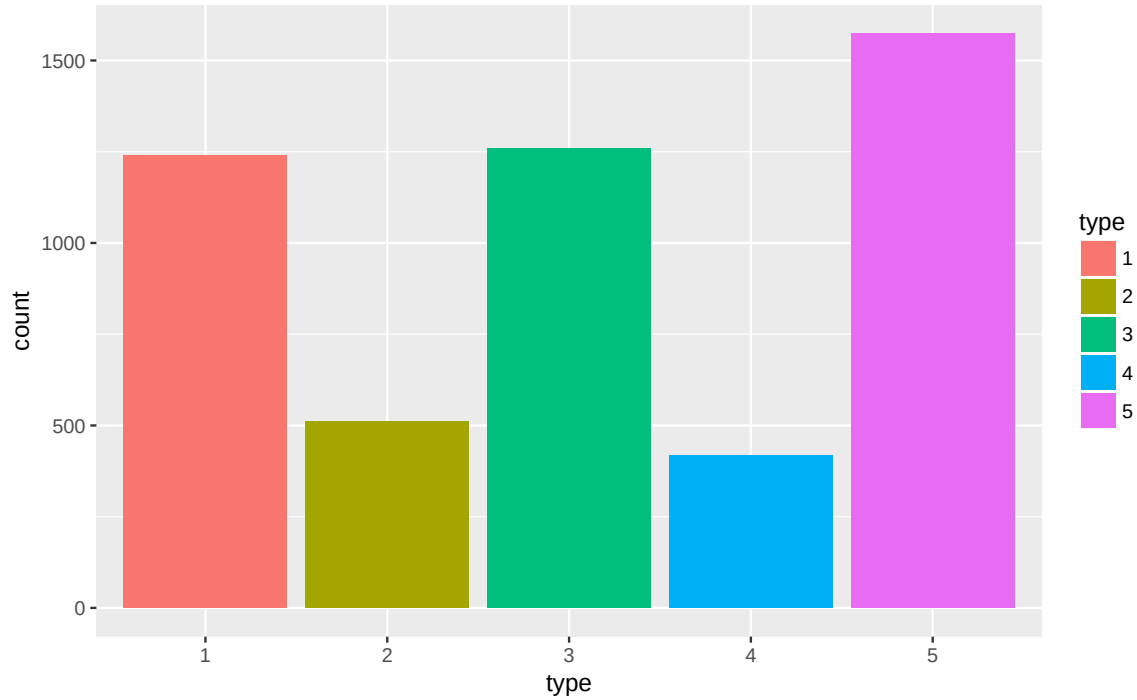

FIG.: Observation pour 5 cellules souches et 1000 cellules différenciées

Pour rappel : la composition attendue dans le sang

B (type 1)	M (type 2)	N (type 3)	T (type 4)	NK (type 5)
0.266	0.114	0.27	0.076	0.274

Données : $n = 1000$ par cellule souche

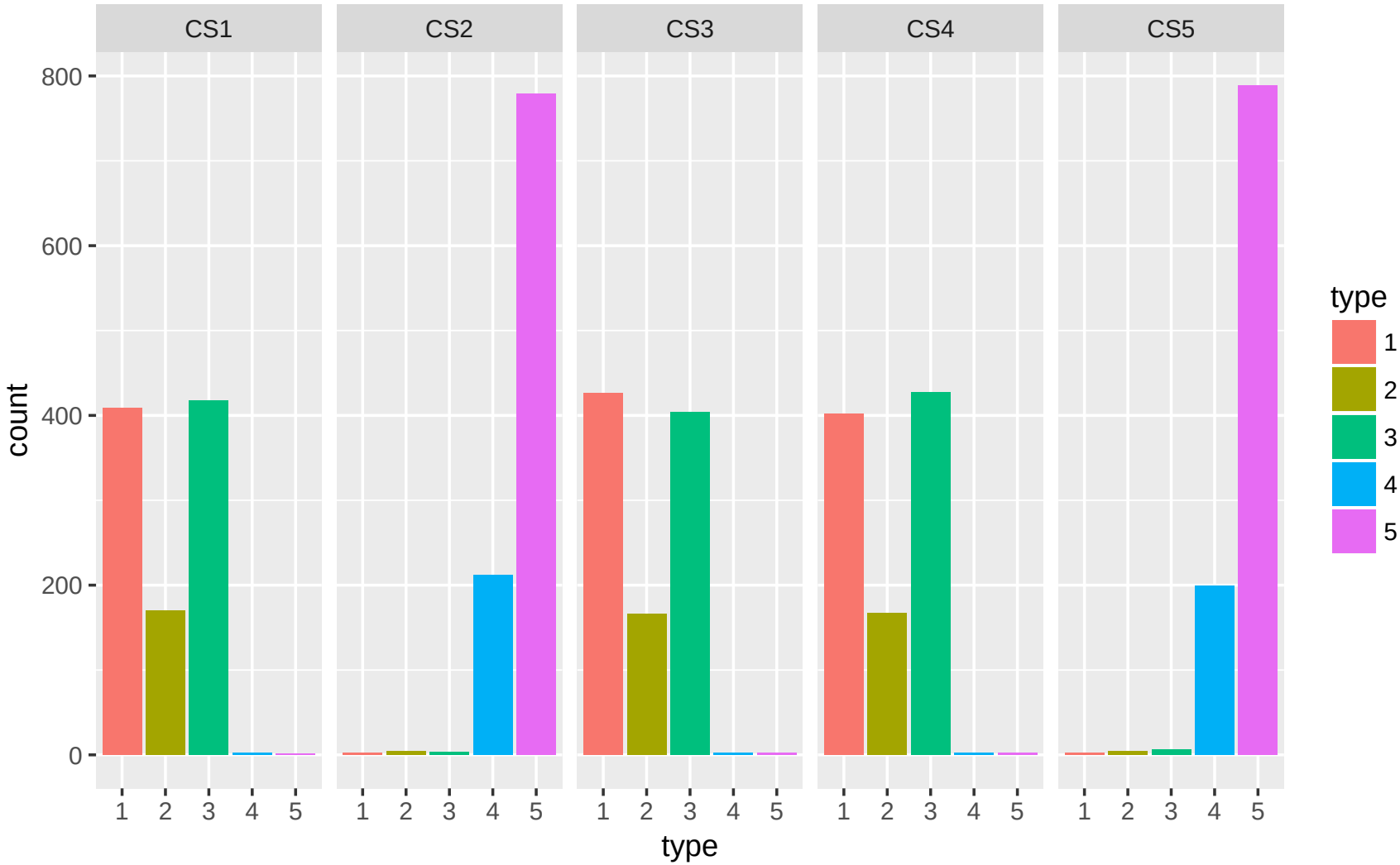

FIG.: Observation pour 5 cellules souches et 1000 cellules différenciées

Un pas en arrière : la loi multinomiale

Une cellule souche produit n cellules différenciées

B B M N T NK NK B M NM ...

ou bien

B	M	N	T	NK
n_B	n_M	n_N	n_T	n_{NK}

La loi multinomiale $\mathcal{M}(n, \pi_B, \pi_M, \pi_N, \pi_T, \pi_{NK})$ donne alors

$$\begin{aligned}
&\mathbb{P}(N_B = n_B, N_M = n_M, N_N = n_N, N_T = n_T, N_{TK} = n_{NK}) \\
&= \frac{n!}{n_b! \dots n_{NK}!} \pi_B^{n_B} \dots \pi_{NK}^{n_{NK}}
\end{aligned}$$

Pour introduire de l'hétérogénéité

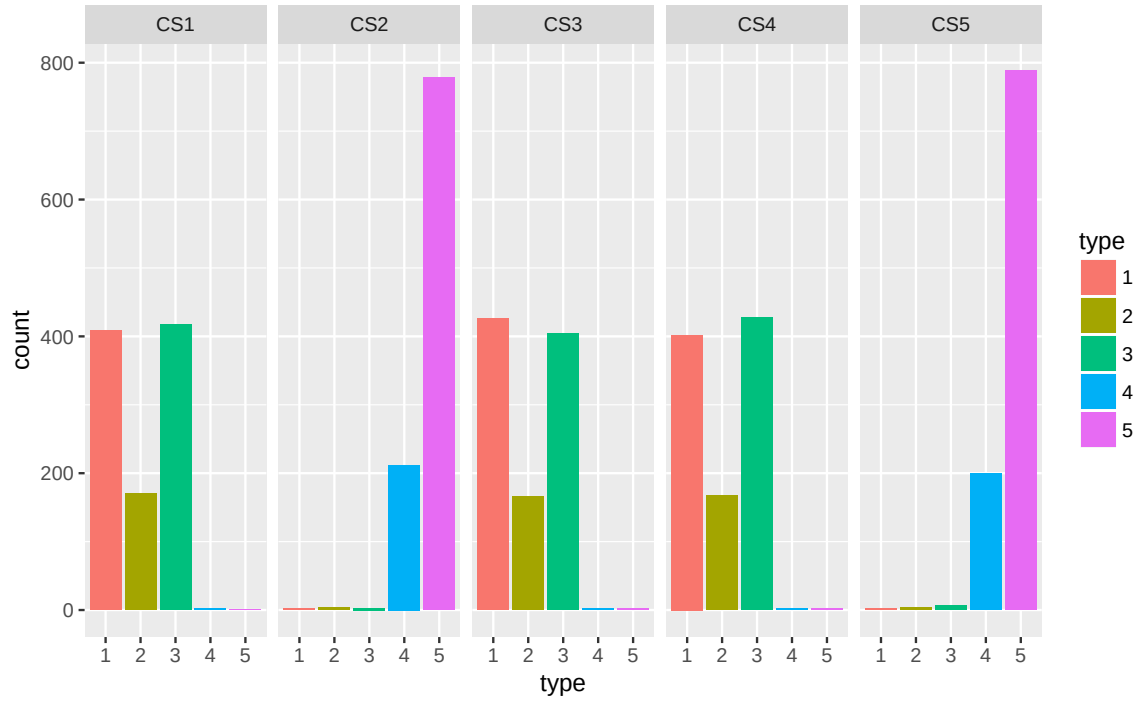

FIG.: Observation pour 5 cellules souches et 1000 cellules différenciées

Il faudrait que toutes les cellules souches n'aient **pas la même loi** $\mathcal{M}(n, \pi_B, \pi_M, \pi_N, \pi_T, \pi_{NK})$.

Un modèle pour l'hétérogénéité : PLSA

- On suppose qu'il existe une **variable non-observée** (variable latente), appelée topic,
- telle que conditionnellement à cette variable, on connaît la loi de probabilité de production des différents types de cellules différenciées, c'est-à-dire

$$\pi^{\mathrm{topic}} = (\pi_B^{\mathrm{topic}}, \pi_M^{\mathrm{topic}}, \pi_N^{\mathrm{topic}}, \pi_T^{\mathrm{topic}}, \pi_{NK}^{\mathrm{topic}}).$$

PSLA $=$ analyse sémantique latente probabiliste [Hof99], issu de la littérature scientifique sur le "text mining" (fouille de données textuelles)

Dans l'exemple

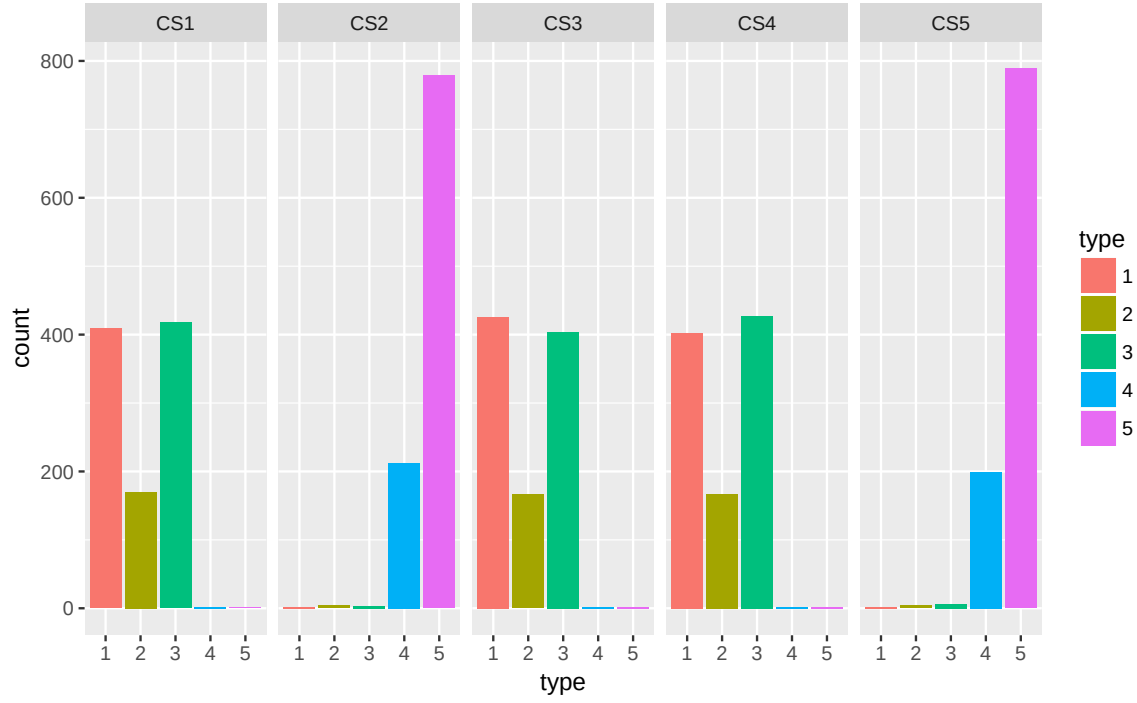

FIG.: Observation pour 5 cellules souches et 1000 cellules différenciées

La variable latente topic vaut 1 ou 2.

Dans l'exemple (suite)

- si une cellule souche est de $\mathrm{topic}\ 1$, elle produit

	B	M	N	T	NK
$\pi^1 =$	0.41	0.17	0.41	0.00	0.00

- si une cellule souche est de $\mathrm{topic}\ 2$, elle produit

	B	M	N	T	NK
$\pi^2 =$	0.00	0.00	0.00	0.22	0.78

On a choisi

$$\mathbb{P}(\mathrm{topic} = 1) = 2/5$$
$$\mathbb{P}(\mathrm{topic} = 2) = 3/5$$

Loi de Bayes

Pour K "topics" avec

- $\mathbb{P}(\text{topic} = 1), \ldots, \mathbb{P}(\text{topic} = K)$
- et $\pi^{\text{topic=k}} = (\pi_B^{\text{topic=k}}, \pi_M^{\text{topic=k}}, \pi_N^{\text{topic=k}}, \pi_T^{\text{topic=k}}, \pi_{NK}^{\text{topic=k}})$

on utilise la loi de Bayes pour calculer

$$\begin{aligned}
&\mathbb{P}\big(\text{topic} = \text{k} | N_B = n_B, \ldots, N_{TK} = n_{NK}\big) \\
&= \frac{\mathbb{P}(\text{topic} = \text{k})\mathbb{P}(N_B = n_B, \ldots, N_{TK} = n_{NK} | \text{topic} = \text{k}))}{\sum_{k'} \mathbb{P}(\text{topic} = \text{k}')\mathbb{P}(N_B = n_B, \ldots, N_{TK} = n_{NK} | \text{topic} = \text{k}'))}
\end{aligned}$$

La statistique fait le reste...

D'où viennent les données ?

- Essai de **thérapie génique**
- pour le syndrome de Wiskott-Aldrich (déficit immunitaire dû à une mutation génétique)
- développé par le laboratoire de Mmes Cavazzana et André-Schmutz (institut Imagine)

Les résultats

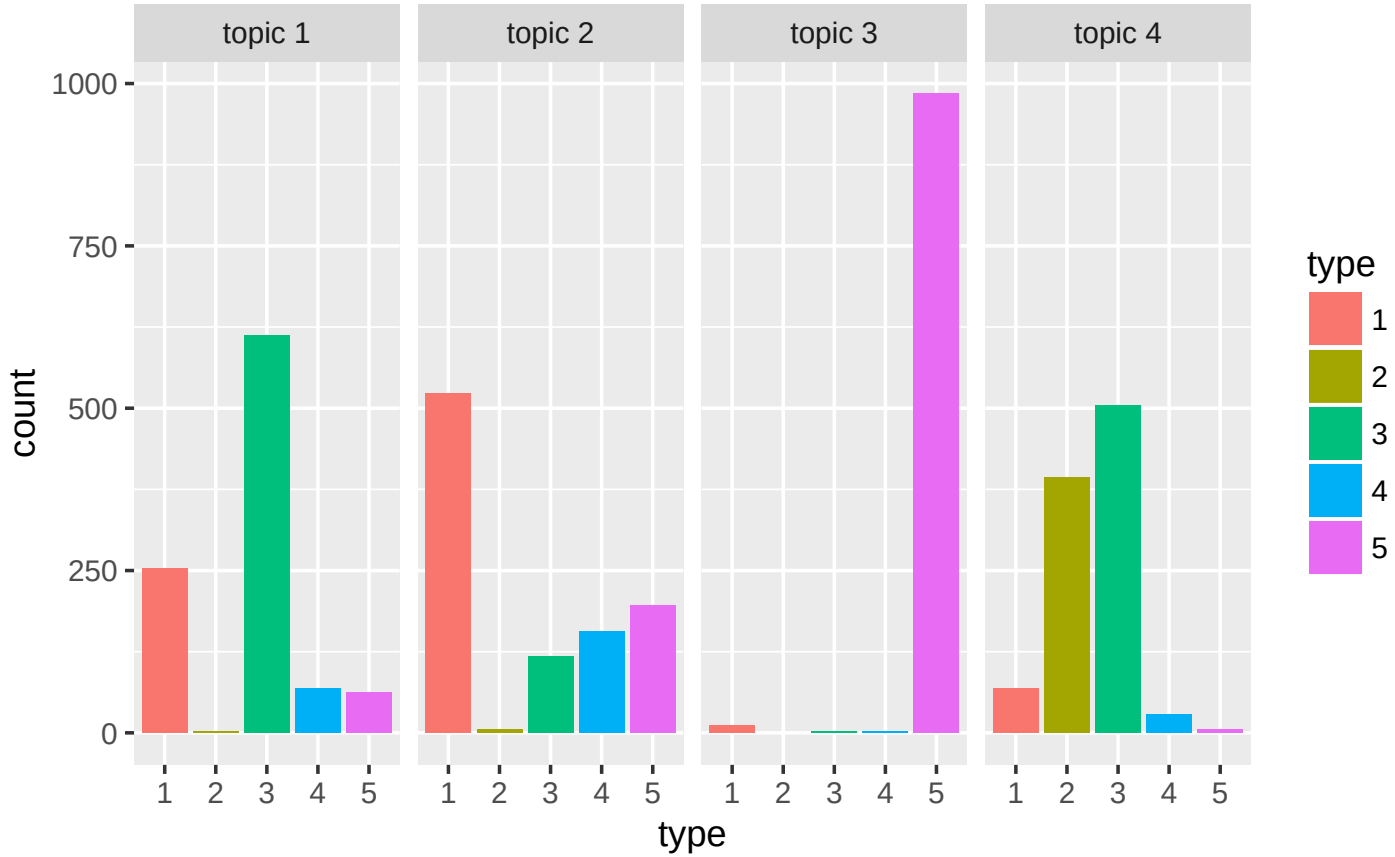

FIG.: Nos résultats

References I


Andrew J Becker, Ernest A McCulloch, and James E Till, *Cytological demonstration of the clonal nature of spleen colonies derived from transplanted mouse marrow cells*.


Brad Dykstra, David Kent, Michelle Bowie, Lindsay McCaffrey, Melisa Hamilton, Kristin Lyons, Shang-Jung Lee, Ryan Brinkman, and Connie Eaves, *Long-term propagation of distinct hematopoietic differentiation programs in vivo*, Cell stem cell **1** (2007), no. 2, 218–229.


Thomas Hofmann, *Probabilistic latent semantic indexing*, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 1999, pp. 50–57.


Wikipedia, *Cellule souche — Wikipedia, the free encyclopedia*, 2017, [Online ; accessed 2-February-2017].

References II

 _____, *Système immunitaire inné— Wikipedia, the free encyclopedia*, 2017, [Online ; accessed 2-February-2017].