Devoir I : Concentration de la distribution des prénoms dans le temps et l'espace

Objectifs

Attendu : une fichier au format `Rmarkdown` (`.Rmd`) ou `Quarto` (`.qmd`), compilable en format `html`.

Dans ce fichier, on trouvera le code nécessaire à la génération des graphiques et des tables correspondants aux questions ci-dessous.

Mesures de l'inégalité et de la diversité

Les mesures de l'inégalité d'une distribution ont été élaborées pour étudier des questions posées en:

- Économie (distribution des revenus, des patrimoines, politiques de redistribution, …)
- Biodiversité (distribution des espèces vivant sur un territoire, …)
- Politique,
- …

> **Tip**
>
> Le package `R` nommé `ineq` met à disposition un certain nombre de concepts élaborés depuis plus d'un siècle.

On peut aussi utiliser ces mesures de l'inégalité/de la concentration pour étudier la distribution des prénoms donnés chaque année dans une société (un pays, une région, …)

Schématiquement, une distribution sur une population de n individus, est un vecteur de longueur n à coefficients positifs, que l'on notera x. Dans le cas qui nous intéresse, si n prénoms distincts ont été utilisés durant une année, x_i sera le nombre de fois où le prénom i aura été attribué dans l'année, $p_i \overset{\text{def}}{=} x_i / \left(\sum_{j=1}^{n} x_j \right)$, représentera la part des nouveaux-nés ayant reçu le prénom i (la *popularité* du prénom i).

Quand on s'intéresse au caractère plus ou moins inégalitaire d'une distribution, l'important n'est pas qui reçoit quoi, mais quelle quantité reçoivent les mieux lotis ou les moins bien lotis. Si une distribution y sur les n prénoms peut être obtenue à partir de x en échangeant les rôles de quelques prénoms (par exemple en renommant les `Pierre` en `Paul` et vice versa), y est ni plus ni moins inégalitaire que x. Une bonne mesure de l'inégalité des distributions doit être *invariante par permutation*.

Dans la suite, on suppose les vecteurs représentants les permutations triés par ordre croissant. Si x et y représentent deux distributions sur n individus (prénoms pour nous). On dira que $x \preccurlyeq y$ (y majore x) si pour tout $i \leq n$

$$\sum_{j \leq i} p_j = \frac{\sum_{j \leq i} x_j}{\sum_{j \leq n} x_j} \geq \frac{\sum_{j \leq i} y_j}{\sum_{j \leq n} y_j} = \sum_{j \leq i} q_j$$

autrement dit, si pour tout $i \leq n$, la part de richesse allouée au i moins bien lotis dans x est plus grande que dans y.

Deux distributions de richesse sur un même ensemble ne sont pas toujours comparables au sens de l'ordre $\preccurlyeq$: il se peut que l'on n'ait ni $x \preccurlyeq y$, ni $y \preccurlyeq x$. Il s'agit d'un *ordre partiel*.

La courbe de Lorenz L_x d'un vecteur x est définie par

$$L_x(i/n) = \frac{\sum_{j \leq i} x_j}{\sum_{j \leq n} x_j} \qquad \text{pour } 1 \leq i \leq n$$

et

$$L_x(t) = L_x(i/n) \qquad \text{pour } \frac{i-1}{n} < t \leq \frac{i}{n}$$

Dessiner les courbes de Lorenz des vecteurs x et y permet de visualiser la relation $\preccurlyeq$: si $x \preccurlyeq y$ alors $L_x(t) \geq L_y(t)$ pour tout $t \in]0, 1[$.

Les courbes de Lorenz étant des fonctions de $]0, 1[$ dans $[0, 1]$, on peut superposer sur un même graphique des courbes de Lorenz définies par des vecteurs de longueur différentes. Pour nous, cela permet de comparer les distributions de prénoms des deux sexes dans un même pays durant une même année, de comparer des années différentes, *etc.*

1. Si ϕ doit être utilisée comme une mesure, un indice d'inégalité alors

   $$x \preccurlyeq y \Longrightarrow \phi(x) \leq \phi(y)$$

   soit ϕ est *Schur-convexe*.

2. $\phi(ax) = \phi(x)$ (pour $a > 0$) (invariance par changement d'échelle)

3. L'extremum est atteint lorsque toutes les composantes de x sont égales.

Sources: Section F. de *Measuring Inequality and Diversity* dans *Inequalities: Theory of Majorization and its Applications* Marshall et Olkin. Springer-Verlag.

> **Question**
>
> Calculer pour chaque année, sexe et pays les indicateurs suivants de la dispersion/concentration de la distribution des prénoms

1. Indice de Gini (vu en cours) $1 - 2\int_0^1 L_p(z)\mathrm{d}z$
2. Entropie de Shannon $(p_i)_{i\leq N} \mapsto \sum_{i=1}^N p_i \log_2 p_i$
3. Entropie de Rényi (ordre 2) $(p_i)_{i\leq N} \mapsto -\log_2\left(\sum_{i=1}^N p_i^2\right)$ (voir aussi mesure de diversité de Simpson)
4. Majorité minimale d'Alker: $\inf\{z : L_x(z) \geq 1/2\}$
5. Part du dernier décile: pour $\alpha = 10\%$, $p \mapsto 1 - L_p(1-\alpha)$
6. Indice(s) d'Atkinson $1 - N\left(\frac{1}{N}\sum_{i=1}^N p_i^{1-a}\right)^{1/(1-a)}$ pour $a \in (0,1)$, choisir α.

Le résultat sera conservé dans une table où chaque ligne correspondra à un pays, une année, un sexe (la clé) avec en plus une colonne par indicateur.

i Question

Tracer les graphes de l'évolution de ces indicateurs de la dispersion/concentration de la distribution. Utiliser le mécanisme des *facettes* pour juxtaposer les graphes correspondants aux quatre couples (Pays, Sexe). Pour chaque (Pays, Sexe), superposez les graphes des indicateurs en fonction du temps.

Ajustement à une loi de Zipf

Une distribution de Zipf est classiquement une loi définie à partir le nombre d'occurrences des mots dans un texte. une loi sur $\mathbb{N}$. On utiliser cette approche pour étudier la distribution des prénoms donnés une année aux bébés d'un sexe donné: on range les prénoms par popularité décroissante, et on trace le graphe de la popularité en fonction du rang. Pour visualiser, on choisit des échelles logarithmiques pour les deux axes. On appelle ces graphes des diagrammes de Zipf.

i Question

Tracer les digrammes de Zipf, pour les deux sexes, les USA et la France, pour les années 1950, 1990, et 2015.

Profils de popularité

Nous croyons qu'il existe quatre sortes de prénoms. La première catégorie consiste en des prénoms qui ont connu une baisse continuelle de popularité depuis la seconde guerre mondiale. La deuxième catégorie comprend les noms qui ont connu une hausse continuelle de leur popularité durant cette période. La troisième catégorie est constituée de prénoms qui sont devenus

progressivement à la mode et qui sont ensuite retournés dans l'ombre. La quatrième catégorie est constituée de prénoms qui ont décliné puis connu un regain de popularité.

> **Caution**
>
> "Baisse continuelle", "hausse continuelle" ne sont pas des notions formalisées. Pour donner un sens effectif aux quatre catégories, on peut lisser les popularités en calculant des moyennes mobiles, et analyser les variations des series lissées.

> **Question**
>
> Considérez les prénoms qui ont figuré au moins une fois depuis 1948 parmi les 300 prénoms les plus populaires dans leur pays (pour un genre donné). Proposez une classification de ces prénoms en fonction de l'évolution de leur popularité au cours de 70 dernières années.

Barème

Critère	Points	Détails
Orthographe et grammaire	20%	English/Français
Graphiques	25%	Choix des `aesthetics`, `geom`, `scale` …
Style des Graphiques	15%	Titres, légendes, étiquettes …
Manipulations de tables	25%	
Respect DRY	15%	Principe DRY W Wikipedia