DM 1 - M2 Recherche
Un peu de statistiques descriptives

Emeline Schmisser, `emeline.schmisser@math.univ-lille1.fr`, bureau 315

Dans tout le devoir, on n'utilisera aucune des fonctions moyenne ou variance déjà implémentée dans R.

1 Tables de mortalité

1.1 Présentation

Une table de mortalité donne, pour chaque année, les quotients de mortalité q_k. Le quotient de mortalité à un âge mesure la probabilité, pour les personnes survivantes à cet âge, de décéder avant l'âge suivant (source : INSEE).

$$q_k = \mathbb{P}\,(\text{mourir avant } k+1 \text{ ans} |\ \text{on est vivant à l'âge } k)$$

Cette table de mortalité nous montre que pour les enfants nés en 1806, 20% des enfants mourraient avant leur 1er anniversaire, et 2% des enfants qui ont atteint leur 4ème anniversaire sont morts entre leur 4ème et leur 5ème anniversaire.

	Age				
Année	q0	q1	q2	q3	q4
1806	0.2	0.08	0.04	0.02	0.02
1900	0.2	0.03	0.02	0.01	0.01
1930	0.08	0.02	0.006	0.004	0.003
1960	0.03	0.002	0.001	0.0007	0.0005
1996	0.005	0.0004	0.0003	0.0002	0.0001

1. Télécharger les tables `"morts.csv"`,`"mortsf.csv"`et `"mortsg.csv"`et les ouvrir sous R : changer le répertoire de travail et utiliser la commande `"morts<-read.csv2("morts.csv",header=TRUE,row.names=1)"`. Ces trois tableaux de données donnent les quotients de mortalité (de 0 à 104 ans) pour respectivement toute la population, les femmes et les hommes de 1806 à 1996. Certains quotients de mortalité dans cette table sont des prévisions (on ne sait pas combien de personnes nées en 1996 vont mourir avant 100 ans, par exemple).
2. Comprendre ce que font les commandes
   - `"names(morts)"` ?
   - `"morts[1:6,1:20]` ?
   - `"morts$q0"` ?
   - `"q0"` ?
   - `"attach(morts)"` et `"q0"` ?
   - `"morts[1990,]` ? et `"morts["1990",]"` ?
   - `"morts["1990",]"` et `"as.numeric(morts["1990",])"` ?
3. Créer la fonction `ma` de la manière suivante : ma<-function(a) as.numeric(morts[a,]) Que fait cette commande ? (expliquer)

1.2 Fonction de survie

On note $S(k)$ la proportion de la population qui atteint l'âge k. En particulier,

$$S(0) = 1 \quad \text{et} \lim_{k\to\infty} S(k) = 0).$$

Exercice 1 (4 points)

1. Donner la formule qui permet, à partir des quotients de mortalité q_k, de calculer S.
2. Sous R, construire une fonction `S` qui permet de calculer S (c'est à dire le vecteur $(S(0), S(1), \ldots, S(104))$ en fonction des quotients de mortalité.
3. Tracer sur un même graphique les fonctions de survie pour les années 1810, 1890, 1910, 1930, 1960 et 1990 de la table «morts.csv». On utilisera pour cela les fonctions `mae`t `S`. Pour que le graphique soit plus facile à lire, utiliser une couleur différente par année (red, darkorange, orange, green, darkgreen, darkblue par exemple) Commenter.

1.3 Durée moyenne de vie et durée de vie résiduelle

La durée moyenne de vie E est la moyenne de la durée de vie d'un individu. La durée de vie résiduelle à l'âge k E_k est le temps moyen qu'il reste à vivre à un individu qui a atteint l'âge k.

$$E_k(X) = \mathbb{E}(X - k | X \geq k).$$

Exercice 2 (6 points)

1. Expliquer comment on calcule la durée moyenne de vie à partir de la fonction de survie S. Indication : $S(k) = \mathbb{P}(X \geq k)$ si X représente une personne de la population.
2. Sous R, créer une fonction `Ev` pour calculer la durée moyenne de vie en fonction des quotients de mortalité (utiliser la fonction S déjà créée). Calculer l'espérance de vie pour les années 1810, 1890, 1910, 1930, 1960 et 1990.
3. Donner l'expression de la fonction de survie résiduelle F_k en fonction des quotients de mortalité.
4. Comment calcule-t-on la durée de vie résiduelle en fonction de F_k ?
5. Contruire une fonction `Evr` qui calcule le vecteur $E_0, E_1, E_2, \ldots, E_{105}$ en fonction des quotients de mortalité. Indication : utiliser une boucle for.
6. Comparer sur un même graphique les durées de vie résiduelles pour les années 1810, 1890, 1910, 1930, 1960 et 1990.
   Commenter les résultats. Est-ce que les durées de vie résiduelles diminuent toujours ? Pourquoi ?

1.4 Variance

Exercice 3 (5 points)

1. Expliquer comment on calcule, à partir de S, la proportion de personnes qui sont mortes à l'âge k qu'on appelle M_k. Par convention, on suppose $M_0 = 0$.
2. Construire une fonction `Mo` sous R qui calcule le vecteur $M_0, M_1, \ldots, M_{104}$.
3. Construire une fonction `Evbis` qui calcule l'espérance de vie en utilisant Mo. Calculer l'espérance de vie pour les années 1810, 1890, 1910, 1930, 1960 et 1990. Comparer avec les résultats obtenus dans l'exercice précédent.
4. Donner la formule qui permet de calculer l'écart-type de la durée de vie en fonction du vecteur M_k.
5. Construire une fonction `Va` qui calcule l'écart-type de la durée de vie en fonction des quotients de mortalité. Calculer cette variance pour les années 1810, 1890, 1910, 1930, 1960 et 1990.

En réalité, on se sert assez peu de la variance quand on étudie les tables de mortalité. On préfère regarder l'espérance de vie et certains quotients de mortalité particuliers : les quotients de mortalité infantile, par exemple.

1.5 Génération fictive

On peut calculer la durée moyenne de vie pour les générations nées en 1800 ou en 1890 : pratiquement tous leurs représentants sont morts. Mais l'INSEE calcule quand même une espérance de vie. C'est la durée moyenne de vie d'une génération fictive soumise aux conditions de mortalité de l'année. On va comparer la génération fictive de 1890 et les taux réels de mortalité qu'a connu cette population. Pour cette étude, on sépare les hommes et les femmes.

Exercice 4 (4 points)

1. Construire deux vecteurs `mfh` et `mff` qui contiennent les quotients de mortalité correspondant à l'année 1890 (et non aux personnes nées en 1890). Indication : utiliser une boucle for. Les quotients de mortalité des gens nés en 1890 se trouvent sur la ligne 189 du tableau.
2. Expliquer comment, à partir de ces deux vecteurs, on peut calculer la fonction de survie de la génération fictive et l'espérance de vie.
3. Comparer l'espérance de vie de la génération fictive et la durée moyenne de vie.
4. Comparer les fonctions de survie de la génération fictive et de la vraie génération de 1890 (pour les hommes et les femmes). Les résultats sont-ils les mêmes pour les hommes et les femmes ? Expliquer pourquoi.
   Données : INED Définitions : INSEE

2 Inégalités salariales

La courbe de Lorentz et l'indice de Gini sont de bons moyens d'étude d'inégalité. La courbe de Lorenz est la représentation graphique de la fonction qui à la part x des salariés les moins riches associe la part y du revenu total qu'ils perçoivent. Donc, si on note $s_1, s_2, \ldots, s_n$ les salaires (ordonnés), S la somme de tous les salaires de l'entreprise, alors :

$$L(k) = \sum_{i=1}^{k} s(i)/S.$$

L'indice de Gini est 2 fois l'aire entre la courbe de Lorentz et la première bissectrice.

Exercice 5 (11 points)

Charger le fichier `salaires.csv` dans R.

1. Donner les moyennes et les écart-type des salaires par sexe. Faire une décomposition de la variance pour étudier les disparités de salaires entre les hommes et les femmes.
2. Quel(s) test(s) pourrait-on réaliser pour s'assurer que la différence de salaire entre les hommes et les femmes n'est pas dûe au hasard ? On s'intéresse maintenant aux inégalités de répartition des salaires (pour l'ensemble des salariés).
3. Si la société est complètement égalitaire (tout le monde gagne la même somme), comment se caractérise la courbe de Lorentz ? L'indice de Gini ? Même question si la société est totalement inégalitaire (une seule personne possède toutes les richesses).
4. Comment évolue l'indice de Gini si la société devient plus égalitaire ? plus inégalitaire ?
5. Donner la formule de l'indice de Gini en fonction du vecteur s.
6. Tracer sur le même graphique la courbe de Lorentz et la première bissectrice.
7. Calculer l'indice de Gini. Les salaires d'une entreprise sont répartis comme suit :

Proportion	Salaires
10%	1000
30%	1200
20%	1300
15%	1500
10%	1700
10%	1900
5%	2300

8. Donner la formule théorique pour calculer la courbe de Lorentz en fonction des vecteur Salaires s et Proportion p.
9. Donner la formule théorique du coefficient de Gini.
10. Tracer sur un même graphique la courbe de Lorentz et la première bissectrice pour cette entreprise.
11. Calculer le coefficient de Gini.