DEVOIR

M1 Mathématiques Appliquées, Statistique
année 2024-2025
cours de Classification

Ce devoir est à remettre par binôme à la rubrique devoir de la page AMeTICE du cours en y déposant un répertoire nommé vosnoms.zip constitué d'un fichier au format pdf contenant les réponses aux questions et d'un fichier programme contenant les codes `R` demandés. La date limite est fixée au vendredi 28 mars.

Partie 1

Présentation des données

Le jeu de données considéré[1] est relatif aux caractéristiques en matière d'emploi de quinze pays européens. Pour chacun de ces pays, on dispose des variables relatives à l'emploi pour l'année 2009 exprimées en pourcentage et présentées dans la table 1. Le jeu de données est disponible sur Ametice dans la rubrique "Devoir".

Variable	Libellé
TxEmp	Taux d'emploi des individus âgées de 15 à 64 ans
TauxEmpFem	Taux d'emploi des femmes âgées de 15 à 64 ans
TxEmpJeun	Taux d'emploi des individus âgées de 15 à 24 ans
TxEmpSen	Taux d'emploi des individus âgées de 55 à 64 ans
Tpart	Part de salariés à temps partiel parmi les individus en emploi
CDD	Part de salariés en contrats temporaires parmi les individus en emploi
TxCho	Taux de chômage
TxChoLD	Taux de chômage de longue durée en pourcentage de la population active

TABLE 1 – Nom et libellé des variables constituant le jeu de données.
NB : Selon la définition de l'INSEE, le taux d'emploi d'une classe d'individus est calculé en rapportant le nombre d'individus de la classe ayant un emploi au nombre total d'individus dans la classe. Il peut être calculé sur l'ensemble de la population d'un pays, mais on se limite le plus souvent à la population en âge de travailler (généralement définie, en comparaison internationale, comme les personnes âgées de 15 à 64 ans), ou à une sous-catégorie de la population en âge de travailler (femmes de 15 à 64 ans par exemple).
Source : `www.insee.fr/fr/metadonnees/definition/c1332`

Dans le but de construire une typologie des pays quant à leurs caractéristiques en matière d'emploi, on va réaliser une classification ascendante hiérarchique.

1. Donner le vecteur associé au point "Hongrie". Que proposeriez-vous comme distance entre deux individus ? Entre deux classes ?

2. En déduire la valeur de l'inertie du nuage de points.

3. Réaliser la Classification Ascendante Hiérarchique sous R.

4. Quels sont les éléments réunis à la première et à la dernière agrégation de l'algorithme ?

5. Quel est, à l'issue de la sixième étape de l'algorithme, le nombre de classes ayant strictement plus de deux pays ? Vous donnerez le contenu de chacune de ces classes. Quelle est l'inertie intra-classes et l'inertie inter-classes à l'issue de cette étape ?

1. Source : Labor Force Survey, Eurostat

6. Donner les définitions et l'interprétation du R carré et du R carré semi partiel.

7. Représenter la hiérarchie des partitions obtenues sous la forme d'un dendogramme. Que représente la hauteur des branches du dendogramme ? A partir du vecteur "height" fourni par `R`, calculer le vecteur des distances entre chaque agrégation.

8. Choisir une partition en k classes des individus. Justifier ce choix à partir d'indicateurs et de graphiques que vous construirez sous `R`. Donner le contenu de chacune de ces k classes.

9. Calculer les moyennes par classe, puis les valeurs tests associées à la partition en k classes précédente (en utilisant le programme donné dans la fiche 4 question 16). En déduire une interprétation de chaque classe.

Partie 2

Présentation des données

La table 2 est issue du site[2] education.gouv.fr du ministère de l'éducation nationale, de la jeunesse et des sports. Elle présente le nombre d'étudiant.e.s inscrits à l'université en France pour l'année 2020-2021 suivant leurs origines sociales et les filières suivies.

	AgrArtCommChEnt	CadPrIntel	Pinter	Employés	Ouvriers	Retraités-inactifs
Droit sciences politiques	19041	68475	25592	32148	17771	27960
Economie, AES	21279	54497	27658	37731	25970	28790
Arts, lettres, langues, SHS	35988	122705	72565	91446	53805	85196
Sciences	26757	103009	49844	50914	33089	36909
Staps	5252	18114	11388	12674	7126	5328
Santé	18081	96241	26260	21964	12470	24439

TABLE 2 – Nombre d'occurrences des filières universitaires suivies suivant l'origine sociale des étudiants. L'origine `AgArtCommChEnt` signifie Agriculteurs, artisans, commerçants et chefs d'entreprise, l'origine `CadPrIntel` signifie Cadres et professions intellectuelles supérieures, l'origine `Pinter` signifie Professions Intermédiaires.

1. Comment s'appelle le tableau présenté table 2 ? Définissez les variables qui lui sont associées.

2. Quelle distance proposeriez-vous pour calculer la distance entre les filières (justifier). Calculer la distance entre les filières "Sciences" et "Staps".

On s'intéresse désormais plus précisément au choix de la filière **Sciences** selon l'origine sociale de l'étudiant.e.

3. Compléter les cases vides du tableau des effectifs suivant :
4. Calculer les estimations des quantités suivantes :

- La cote de choisir la filière `sciences` contre le fait de choisir une autre filière pour un.e étudiant.e appartenant au groupe professionnel `employés`,
- la filière `sciences` contre le fait de choisir une autre filière pour un.e étudiant.e appartenant au groupe professionnel `ouvriers`,
- Le rapport de cotes (odds-ratio) pour une personne appartenant au groupe professionnel `ouvriers` par rapport à une personne appartenant au groupe professionnel `employés`.

2. `https://www.education.gouv.fr/reperes-et-references-statistiques-2021-308228`

	AgrArtCommChEnt	CadPrIntel	Pinter	Employés	Ouvriers	Retraités-inactifs
Sciences	26757	103009	49844	50914	33089	36909
Autre	99641		163463	195963		171713

TABLE 3 – Nombre d'occurrences de la filière universitaire `Sciences` et de l'ensemble des filières autres que `Sciences` suivies suivant l'origine sociale des étudiants. L'origine `AgArtCommChEnt` signifie Agriculteurs, artisans, commerçants et chefs d'entreprise, l'origine `CadPrIntel` signifie Cadres et professions intellectuelles supérieures, l'origine `ProfInter` signifie Professions Intermédiaires.

- Interpréter les résultats précédents.

5. Définir le modèle de régression logistique permettant de modéliser la probabilité de choisir la filière **sciences** pour un.e étudiant.e en fonction de son origine sociale. On considérera comme profil de référence une personne provenant du groupe professionnel `employés`.

6. Mettre en oeuvre ce modèle sous `R`.

7. Donner l'estimation du maximum de vraisemblance du coefficient du modèle associé à une personne appartenant au groupe professionnel `ouvriers`, à partir des sorties `R`. Retrouver ce résultat à partir des questions précédentes.

8. Donner l'estimation de la variance de l'estimateur de ce coefficient et la valeur de la statistique du test de Wald. Comment est calculée cette statistique ? Que permet-elle de tester (on donnera l'hypothèse nulle et l'hypothèse alternative) ?

9. Quelle est la valeur de la déviance du modèle ? Donner sa formule dans ce cas particulier.

10. Interpréter les sorties du modèle.