Université d'Orléans
Science & Technology
1, rue de Chartres - BP 6759 - 45067 Orléans cedex 2

ANNÉE UNIVERSITAIRE 2021-2022 - Semestre 1
EXAMEN 1ère SESSION

ÉPREUVE

Intitulé : Analyse de données

Date et horaire : 13 Décembre 2021, 9h00

Durée : Projet "Sprint" sur la journée, fin à 18h

Responsable de l'UE : Didier Chauveau

Mention/Spécialité/Parcours : M1 Miage

Documents autorisés : Notes de cours, aide et codes logiciel R vus en TP

Matériels autorisés : Machines personnelles ou salles infos avec R

L'objectif de ce **projet SPRINT** est de vous mettre en situation réelle devant un gros fichier de données issu de l'enquête nationale *Emploi 2001* de l'INSEE, qui a pour objectif l'étude des thèmes suivants: Emploi, population active, chômage, marché du travail, activité professionnelle, durée du travail, précarité, recherche d'emploi, mobilité professionnelle, mobilité sociale et formation. Le fichier pédagogique issu de cette enquête dont vous disposez est un échantillon au dixième des ménages. Il comporte 17 687 individus et 113 variables. Pour chaque question, on précisera les individus, variables et techniques statistiques à utiliser.

Vous êtes donc dans la situation d'un *Data Scientist* qui doit effectuer quelques analyses ciblées à partir des fichiers d'origine (dans leur format d'origine, voir ci-dessous) et des documentations de l'enquête. Les analyses demandées sont standards et ont toutes été vues en cours ou TP, mais vous allez devoir passer une partie du temps de la journée à **vous approprier les données**: définitions des variables; codages, formats, classes dans R etc.... Cela fait partie du travail préliminaire et indispensable pour effectuer des analyses correctes, et obtenir des résultats et graphiques lisibles, et fait donc partie de l'évaluation.

Données et éléments de l'enquête (sur le wiki)

- Les données au format R (746 Ko): `Emploi01.Rdata`
- La liste et la description des variables: `Variables_emploi2001.pdf`
- des codes de prétraitement pour vous aider à démarrer pour chaque Question: `SPRINT_Q1.R`, `SPRINT_Q2.R` et `SPRINT_Q3.R` (voir le sujet).

Précisions sur la structure des données et les documents: Le fichier dont vous disposez contient les données avec le codage tel qu'il arrive après import dans R depuis le fichier disponible en ligne au format d'export de SAS. En particulier, les variables arrivent avec la classe `factor`, même si elles sont en fait parfois numériques, et les éventuels manquants ne sont pas codés `NA`, mais par la chaîne vide `""`. Des indications et exemples de codes de démarrage afin de convertir ces variables correctement vous sont donnés dans le sujet pour chaque question.

Présentation et envoi de votre travail

Le document unique, au format pdf, que vous devez nous envoyer avant 18h par mail à la fin de l'épreuve est un texte de synthèse argumentée, compréhensible par un non-spécialiste et pertinente pour un statisticien, contenant les graphiques référencés et interprétés dans le texte, ainsi que en annexe dans le même document vos programmes R.

→ **Envoi par mails à:**
`didier.chauveau@univ-orleans.fr` et `laurent.delsol@univ-orleans.fr`
avec le sujet "`SPRINT rapport groupe n`", où $n \in \{1, 2, \ldots\}$ est votre numéro de groupe (cf. liste envoyée).

Questions

Pour chaque question, on précise les individus, variables et techniques statistiques à utiliser.

1. Statistiques descriptives: typologie du travail à temps partiel chez les actifs

Pour cette question, vous disposez du code de démarrage `SPRINT_Q1.R` qui effectue la sélection des variables utiles et vous donne des exemples de recodages adéquats (traitement des manquants et variables quantitatives), et dont vous vous inspirerez pour ce projet, afin d'éviter les erreurs élémentaires.

1. Limitez l'étude à la sous-population des actifs selon la variable FI, de 20 à 50 ans selon la variable AGCM. Combien reste-t-il d'individus?
2. Il est parfois nécessaire de recoder les facteurs comprenant de nombreuses modalités pour améliorer lisibilité et pertinence des synthèses:
   (a) Recodez Région (RG) en 7 ou 8 "grandes régions" en perdant le moins possible l'information.
   (b) Recodez de même NR (Nationalité) en 4 "groupes" de nationalités cohérents.
3. Analysez, avec des outils numériques et/ou graphiques du chapitre 1 du cours, les liens entre temps partiel/complet (TP) et les variables âge (AGCM), sexe (S), niveau de diplôme (DDIPL), nationalité (NR), région (RG).
4. Effectuez le ou les recodages appropriés afin de réaliser des tests du χ^2 pour l'hypothèse nulle "pas de lien" entre TP et chacun de ces facteurs. Interprétez les résultats afin de déterminer les facteurs les plus significatifs.

2. ACP des variables quantitatives et qualitatives ordinales

Il y a peu de variables quantitatives dans ces données, mais certaines variables qualitatives ont des relations d'ordre naturelles entre modalités, ce qui permet de les considérer comme quantitatives. Ainsi on peut proposer une ACP des 10 variables: AGCM, SALRED, SALFR, DUHAB, NP, NBCHMEN, PIECES, TU90, ADFE, NEGR, afin d'étudier les éventuels liens entre celles-ci et quelques facteurs qualitatifs déjà vus dans la question précédente. Un code de démarrage `SPRINT_Q2.R`, qui a généré le fichier de données spécifique `EmploiQ2.Rdata` vous est fourni pour cette question; dans ce code:

- les variables utiles ont été sélectionnées, et les qualitatives ordinales converties en numérique
- Les lignes contenant des individus non renseignés (`NA`) pour les variables numériques ont été supprimées

1. Sélectionner la sous-population composée des actifs selon la variable FI.
2. Supprimer les individus correspondants à des salaires extrêmes, supérieurs au quantile d'ordre 99% de la variable SALRED. Combien cela supprime-t-il d'individus? A votre avis, pourquoi faire cela?
3. Réaliser l'ACP sur la table résultat de ces traitements et interpréter les sorties.
4. Représenter sur le plan principal (ψ^1, ψ^2) et le plan factoriel (ψ^3, ψ^4) les barycentres des projetés des individus par modalités des trois facteurs S, DDIPL, RG recodés comme pour la Question 1 (cf. techniques vues en TP).
5. Interprétations et synthèse

3. AFC: liens entre diplôme et région

Dans cette partie on s'intéresse à la sous-population composée des {actifs, chomeurs, étudiants, retraités}, selon la variable FI. Le code `SPRINT_Q3.R` vous donne les instructions pour convertir correctement la table de contingence que vous obtiendrez en `data.frame` nécessaire pour l'AFC. Réalisez une Analyse Factorielle des Correspondances entre les facteurs Diplôme (DDIPL) et région (RG) recodée comme pour la question 1. Faites-en l'interprétation et la synthèse.