Statistiques descriptives et tabulation avec gtsummary

Malick SENE

2025-02-11

# Introduction

Ce document présente une analyse descriptive détaillée des données EHCVM2021 pour le Mali. Grâce aux packages **haven**, **dplyr**, **labelled** et **gtsummary**, nous importons, manipulons et résumons ces données pour en extraire des statistiques descriptives personnalisées.

Pour commencer, nous définissons une liste de packages requis, vérifions si chacun est installé et, si nécessaire, l’installons avec ses dépendances, puis nous les chargeons dans l’environnement.

# Définition de la liste des packages requis  
packages <- c("gtsummary", "dplyr", "readr", "haven")  
  
# Vérification de l'installation et installation des packages manquants  
for (pkg in packages) {  
 if (!requireNamespace(pkg, quietly = TRUE)) {  
 install.packages(pkg, dependencies = TRUE)  
 }  
}  
  
# Chargement des packages dans l'environnement  
for (pkg in packages) {  
 library(pkg, character.only = TRUE)  
}

Ensuite nous utilisons la fonction read\_dta() du package **haven** pour importer le fichier Stata (.dta).

Puis, le package **dplyr** offre des fonctions efficaces pour la manipulation des données.  
Ici, nous utilisons la fonction select() pour extraire uniquement les colonnes d’intérêt, à savoir toit et sol.

Enfin, le package **gtsummary** fournit des outils pour créer des tableaux de statistiques descriptives élégants et personnalisables.  
La fonction tbl\_summary() génère un tableau récapitulatif des variables sélectionnées.

# Importation du fichier Stata à l'aide de la fonction read\_dta() du package haven  
data <- haven::read\_dta("../data/ehcvm\_menage\_mli2021.dta")  
  
# Sélection des colonnes 'toit' et 'sol' avec dplyr::select()   
# et génération d'un tableau résumé avec gtsummary::tbl\_summary()  
data %>%  
 dplyr::select(toit, sol) %>%  
 gtsummary::tbl\_summary()

| **Characteristic** | **N = 6,143***1* |
| --- | --- |
| toit en materiaux definitifs |  |
| 0 | 2,700 (44%) |
| 1 | 3,443 (56%) |
| Sol en materiaux definitifs |  |
| 0 | 3,274 (53%) |
| 1 | 2,869 (47%) |
| *1*n (%) | |

Ici, nous convertissons les variables en facteurs en utilisant la fonction to\_factor() du package **labelled**.  
Cette conversion permet d’utiliser les étiquettes associées aux variables dans le tableau récapitulatif, ce qui améliore la lisibilité des résultats.

Ensuite, nous utilisons la fonction select() du package **dplyr** pour extraire les colonnes toit et sol du jeu de données.  
Enfin, la fonction tbl\_summary() du package **gtsummary** est employée pour générer un tableau de statistiques descriptives pour ces variables.

# Conversion des variables en facteurs à l'aide de labelled::to\_factor(),  
# extraction des colonnes 'toit' et 'sol' avec dplyr::select()  
# et création d'un tableau résumé avec gtsummary::tbl\_summary()  
data %>%   
 labelled::to\_factor() %>%   
 dplyr::select(toit, sol) %>%   
 gtsummary::tbl\_summary()

| **Characteristic** | **N = 6,143***1* |
| --- | --- |
| toit en materiaux definitifs |  |
| Non | 2,700 (44%) |
| Oui | 3,443 (56%) |
| Sol en materiaux definitifs |  |
| Non | 3,274 (53%) |
| Oui | 2,869 (47%) |
| *1*n (%) | |

Le code suivant réalise les opérations suivantes :

1. Conversion des variables en facteurs avec labelled::to\_factor(), ce qui permet d’utiliser les étiquettes intégrées dans les données.
2. Sélection des colonnes toit et sol via select() du package **dplyr**.
3. Génération d’un tableau résumé à l’aide de tbl\_summary() du package **gtsummary**, en précisant des libellés personnalisés pour améliorer la lisibilité :
   * La colonne toit est renommée “Toit en materiaux définitif”.
   * La colonne sol est renommée “Sol en materiaux définitif”.

data %>%   
 labelled::to\_factor() %>%   
 dplyr::select(toit, sol) %>%   
 gtsummary::tbl\_summary(  
 label = list(  
 toit ~ "Toit en materiaux définitif",  
 sol ~ "Sol en materiaux définitif"  
 )  
 )

| **Characteristic** | **N = 6,143***1* |
| --- | --- |
| Toit en materiaux définitif |  |
| Non | 2,700 (44%) |
| Oui | 3,443 (56%) |
| Sol en materiaux définitif |  |
| Non | 3,274 (53%) |
| Oui | 2,869 (47%) |
| *1*n (%) | |

Ici, la fonction modify\_header() permet de modifier l’en-tête du tableau pour que la colonne des libellés affiche “Caractéritiques logement”.

library(flextable)  
library(officer)  
library(labelled)  
  
data %>%   
 labelled::to\_factor() %>%   
 dplyr::select(toit, sol) %>%   
 gtsummary::tbl\_summary(  
 label = list(  
 toit ~ "Toit en materiaux définitif",   
 sol ~ "Sol en materiaux définitif"  
 )  
 ) %>%   
 gtsummary::modify\_header(label = "Caractéritiques logement")

| Caractéritiques logement | **N = 6,143***1* |
| --- | --- |
| Toit en materiaux définitif |  |
| Non | 2,700 (44%) |
| Oui | 3,443 (56%) |
| Sol en materiaux définitif |  |
| Non | 3,274 (53%) |
| Oui | 2,869 (47%) |
| *1*n (%) | |

tbl <- data %>%   
 to\_factor() %>%   
 select(toit, sol) %>%   
 tbl\_summary(  
 label = list(  
 toit ~ "Toit en matériaux définitif",   
 sol ~ "Sol en matériaux définitif"  
 )  
 ) %>%   
 modify\_header(label = "Caractéristiques logement")  
  
ft <- as\_flex\_table(tbl)  
ft

| Caractéristiques logement | **N = 6,143**1 |
| --- | --- |
| Toit en matériaux définitif |  |
| Non | 2,700 (44%) |
| Oui | 3,443 (56%) |
| Sol en matériaux définitif |  |
| Non | 3,274 (53%) |
| Oui | 2,869 (47%) |
| 1n (%) | |

Maintenant nous extrayons plusieurs colonnes du jeu de données (toit, sol, superf, grosrum et petitrum) grâce à la fonction select() du package **dplyr**. La fonction tbl\_summary() du package **gtsummary** génère un tableau de statistiques descriptives pour ces variables, avec des libellés personnalisés pour améliorer la lisibilité :

* toit est renommé “Toit en materiaux définitif”
* sol est renommé “Sol en materiaux définitif”
* superf est renommé “Superficie cultivée”
* grosrum est renommé “gros rumunant”
* petitrum est renommé “petit ruminants”

Pour les variables numériques superf, grosrum et petitrum, nous affichons la moyenne et l’écart-type en utilisant l’argument statistic. L’argument digits permet de fixer le format numérique (ici aucune décimale). Enfin, nous modifions l’en-tête du tableau pour que la colonne des libellés affiche “Caractéritiques logement” à l’aide de modify\_header().Nous résumons donc ici l’ensemble des manipulations effectuées sur cette base avec les commentaires explicatifs.

# Conversion en facteurs, sélection des colonnes et création d'un tableau résumé personnalisé  
data %>%   
 labelled::to\_factor() %>% # Conversion des variables en facteurs selon leurs étiquettes  
 dplyr::select(toit, sol, superf, grosrum, petitrum) %>% # Extraction des colonnes d'intérêt  
 gtsummary::tbl\_summary(  
 label = list(  
 toit ~ "Toit en materiaux définitif", # Libellés personnalisés  
 sol ~ "Sol en materiaux définitif",   
 superf ~ "Superficie cultivée",   
 grosrum ~ "gros rumunant",   
 petitrum~ "petit ruminants"   
 ),  
 statistic = list( # Afficher les statistiques  
 superf ~ "{mean} ({sd})",   
 grosrum ~ "{mean} ({sd})",   
 petitrum~ "{mean} ({sd})"   
 ),  
 digits = dplyr::everything() ~ c(0, 0, 0, 0, 0, 0) # Formatage numérique sans décimales  
 ) %>%   
 gtsummary::modify\_header(label = "Caractéritiques logement") # Modification de l'en-tête du tableau

| Caractéritiques logement | **N = 6,143***1* |
| --- | --- |
| Toit en materiaux définitif |  |
| Non | 2,700 (44%) |
| Oui | 3,443 (56%) |
| Sol en materiaux définitif |  |
| Non | 3,274 (53%) |
| Oui | 2,869 (47%) |
| Superficie cultivée | 5 (5) |
| Unknown | 3,286 |
| gros rumunant | 8 (17) |
| Unknown | 2,676 |
| petit ruminants | 13 (23) |
| Unknown | 2,676 |
| *1*n (%); Mean (SD) | |

Ici, nous importons le fichier welfare. Ensuite, nous convertissons les variables en facteurs avec la fonction to\_factor() du package **labelled**, ce qui permet d’utiliser les étiquettes associées aux variables dans le tableau récapitulatif.

Nous utilisons ensuite la fonction select() du package **dplyr** pour extraire les colonnes d’intérêt :  
- hgender (Genre du chef de ménage)  
- hage (Âge du chef de ménage)  
- hmstat (Statut matrimonial)  
- heduc (Niveau d’éducation)  
- hdiploma (Diplômes)

La fonction tbl\_summary() du package **gtsummary** génère un tableau de statistiques descriptives pour ces variables, en personnalisant les libellés pour améliorer la lisibilité :  
- hgender est renommé en “Genre CM”  
- hage est renommé en “Age CM”  
- hmstat est renommé en “Statut matrimonial”  
- heduc est renommé en “Education CM”  
- hdiploma est renommé en “Diplomes CM”

Pour la variable hage, nous affichons la moyenne et l’écart-type en utilisant l’argument statistic.  
L’argument digits fixe le format numérique (ici, aucune décimale).  
Nous affichons systématiquement les valeurs manquantes grâce aux paramètres missing = "always" et missing\_text = "NA".  
Enfin, nous modifions l’en-tête du tableau pour que la colonne des libellés affiche “Caractéritiques Chef de ménage” en utilisant la fonction modify\_header().

# Importation du fichier Stata concernant le bien-être  
wf <- haven::read\_dta("../data/ehcvm\_welfare\_mli2021.dta")  
  
# Conversion des variables en facteurs, extraction des colonnes d'intérêt,  
# création d'un tableau résumé avec libellés personnalisés et formatage des statistiques,  
# affichage des valeurs manquantes et modification de l'en-tête du tableau  
wf %>%   
 labelled::to\_factor() %>%   
 dplyr::select(hgender, hage, hmstat, heduc, hdiploma) %>%   
 gtsummary::tbl\_summary(  
 label = list(  
 hgender ~ "Genre CM",  
 hage ~ "Age CM",  
 hmstat ~ "Statut matrimonial",  
 heduc ~ "Education CM",  
 hdiploma ~ "Diplomes CM"  
 ),  
 statistic = list(  
 hage ~ "{mean} ({sd})" # Affichage de la moyenne et de l'écart-type pour 'hage'  
 ),  
 digits = dplyr::everything() ~ c(0, 0, 0, 0, 0), # Format numérique sans décimales  
 missing = "always", # Affiche systématiquement les valeurs manquantes  
 missing\_text = "NA" # Remplace les valeurs manquantes par "NA"  
 ) %>%   
 gtsummary::modify\_header(label = "Caractéritiques Chef de ménage")

| Caractéritiques Chef de ménage | **N = 6,143***1* |
| --- | --- |
| Genre CM |  |
| Masculin | 5,438 (89%) |
| Féminin | 705 (11%) |
| NA | 0 |
| Age CM | 50 (14) |
| NA | 0 |
| Statut matrimonial |  |
| Célibataire | 85 (1%) |
| Marié(e) monogame | 4,149 (68%) |
| Marié(e) polygame | 1,292 (21%) |
| Union libre | 0 (0%) |
| Veuf(ve) | 535 (9%) |
| Divorcé(e) | 49 (1%) |
| séparé(e) | 33 (1%) |
| NA | 0 |
| Education CM |  |
| Aucun | 4,249 (69%) |
| Maternelle | 4 (0%) |
| Fondamental 1 | 697 (11%) |
| Fondamental 2 | 412 (7%) |
| Secondaire General | 192 (3%) |
| Secondaire Technique et Professionnel | 242 (4%) |
| Superieur | 347 (6%) |
| NA | 0 |
| Diplomes CM |  |
| Aucun | 5,009 (82%) |
| Certificat d'Etudes Primaires (CEP) | 278 (5%) |
| Diplôme d'Etudes Fondamentales (DEF/BEPC) | 218 (4%) |
| Certificat d'Aptitude Professionnelle (CAP) | 85 (1%) |
| Brevet de Technicien (BT) | 166 (3%) |
| Baccalauréat (BAC) | 52 (1%) |
| DEUG, DUT, BTS,IFM | 67 (1%) |
| Licence | 49 (1%) |
| Maitrise | 139 (2%) |
| Master/DEA/DESS | 59 (1%) |
| Doctorat/Phd | 21 (0%) |
| NA | 0 |
| *1*n (%); Mean (SD) | |