RÉPUBLIQUE DU SÉNÉGAL

Un Peuple - Un But - Une Foi

MINISTÈRE DE L'ÉCONOMIE
DU PLAN ET DE LA COOPÉRATION

ANSD

Agence nationale de la Statistique et de la Démographie (ANSD)

ensae

Ecole nationale de la Statistique et de l'Analyse économique Pierre Ndiaye (ENSAE)

Semestre 2 : Projet statistique sous R

Statistiques descriptives avec gtsummary

Rédigé par :
Khadidiatou Diakhaté
Elève en ISEP3

Sous la supervision de :
M. Aboubacre HEMA
Research Analyst

Année scolaire : 2024/2025

Installation des packages si nécessaire

```
packages <- c("haven","dplyr","gtsummary","labelled")


for (package in packages) {
  if (!requireNamespace(package, quietly = TRUE)) {    # Vérifie si le package n'est pa
    install.packages(package)
  }
  library(package, character.only = TRUE) # nom du package en nom ou chaine de caractè
}
```

Exportation des bases

```
library(haven)
ehcvm_menage <- read_dta("donnees/ehcvm_menage_gnb2021.dta") # qui concerne les caract
ehcvm_welfare <- read_dta("donnees/ehcvm_welfare_gnb2021.dta") # qui concerne le bien-
#Visualisons quelques observations de ces bases
head(ehcvm_menage)
```

```
## # A tibble: 6 x 38
##   country    hhid  year grappe menage vague logem    mur  toit   sol eauboi_ss
##   <chr>     <dbl> <dbl>  <dbl>  <dbl> <dbl> <dbl+l> <dbl> <dbl> <dbl>    <dbl>
## 1 GNB    11122701  2021 111227      1     1 2 [Pro~     1     0     0        1
## 2 GNB    11122702  2021 111227      2     1 2 [Pro~     0     1     0        0
## 3 GNB    11122703  2021 111227      3     1 2 [Pro~     0     1     0        0
## 4 GNB    11122704  2021 111227      4     1 2 [Pro~     0     0     0        0
## 5 GNB    11122705  2021 111227      5     1 2 [Pro~     1     0     0        1
## 6 GNB    11122708  2021 111227      8     1 2 [Pro~     0     1     1        1
## # i 27 more variables: eauboi_sp <dbl>, elec_ac <dbl>, elec_ur <dbl>,
## #   elec_ua <dbl>, ordure <dbl>, toilet <dbl>, eva_toi <dbl>, eva_eau <dbl>,
## #   tv <dbl>, fer <dbl>, frigo <dbl>, cuisin <dbl>, ordin <dbl>, decod <dbl>,
## #   car <dbl>, superf <dbl>, grosrum <dbl>, petitrum <dbl>, porc <dbl>,
## #   lapin <dbl>, volail <dbl>, sh_id_demo <dbl>, sh_co_natu <dbl>,
## #   sh_co_eco <dbl>, sh_id_eco <dbl>, sh_co_vio <dbl>, sh_co_oth <dbl>
```

```
head(ehcvm_welfare)
```

```
## # A tibble: 6 x 43
##   grappe menage country  year     hhid vague month      zae    region  milieu
##    <dbl>  <dbl> <chr>   <dbl>    <dbl> <dbl> <date>    <dbl+l> <dbl+l> <dbl+l>
## 1 111227     11 GNB     2021 111227011     1 2021-12-01 7 [Zon~ 1 [Tom~ 2 [Rur~
## 2 111227     46 GNB     2021 111227046     1 2021-12-01 7 [Zon~ 1 [Tom~ 2 [Rur~
## 3 111227      2 GNB     2021  11122702     1 2021-12-01 7 [Zon~ 1 [Tom~ 2 [Rur~
## 4 111227     13 GNB     2021 111227013     1 2021-12-01 7 [Zon~ 1 [Tom~ 2 [Rur~
## 5 111227      3 GNB     2021  11122703     1 2021-12-01 7 [Zon~ 1 [Tom~ 2 [Rur~
## 6 111227     10 GNB     2021 111227010     1 2021-12-01 7 [Zon~ 1 [Tom~ 2 [Rur~
```

Characteristic	N = 5,351[1]
Ocupaçao de alojamento	
1	1,263 (24%)
2	3,095 (58%)
3	574 (11%)
4	419 (7.8%)
Parede em materiais definitivos	1,619 (30%)
Tecto em materiais definitivos	4,586 (86%)
Solo em materiais definitivos	2,737 (51%)

[1] n (%)

```
## # i 33 more variables: hhweight <dbl>, hhsize <dbl>, eqadu1 <dbl>,
## #   eqadu2 <dbl>, hgender <dbl+lbl>, hage <dbl>, hmstat <dbl+lbl>,
## #  hreligion <dbl+lbl>, hnation <dbl+lbl>, hethnie <dbl+lbl>, halfa <dbl+lbl>,
## #  halfa2 <dbl+lbl>, heduc <dbl+lbl>, hdiploma <dbl+lbl>, hhandig <dbl+lbl>,
## #   hactiv7j <dbl+lbl>, hactiv12m <dbl+lbl>, hbranch <dbl+lbl>,
## #  hsectins <dbl+lbl>, hcsp <dbl+lbl>, dali <dbl>, dnal <dbl>, dtot <dbl>,
## #   pcexp <dbl>, zref <dbl>, def_spa <dbl>, def_temp <dbl>, ...
```

Réalisation pas-à pas d'un tableau avec le package gt-summary

Avec la base ehcvm_menage, nous allons présenter quelques statistiques descriptives des variables logem(logement), toit(type de toit), mur(type de mur) et sol(type de sol). Utilisons **tbl_summary** pour cela :

```
library(gtsummary)
ehcvm_menage %>% select(logem,mur,toit,sol) %>% tbl_summary()
```

```
## ! Column(s) "logem" are class "haven_labelled".
## i This is an intermediate datastructure not meant for analysis.
## i Convert columns with `haven::as_factor()`, `labelled::to_factor()`,
##   `labelled::unlabelled()`, and `unclass()`. Failure to convert may have
##   unintended consequences or result in error.
## <https://haven.tidyverse.org/articles/semantics.html>
## <https://larmarange.github.io/labelled/articles/intro_labelled.html#unlabelled>
```

Cependant, avec ce tableau, on ne voit pas les labels des modalités des variables. Pour les afficher, on va utiliser la fonction **to_factor() du package labelled**.

```
library(labelled)
ehcvm_menage %>% labelled::to_factor() %>%select(logem,mur,toit,sol) %>% tbl_summary()
```

Les noms des variables n'étant pas trop explicites, on peut les reformuler avec la commande **label(list(…))** de tbl_summary

Characteristic	N = 5,351[1]
Ocupaçao de alojamento	
Proprietário com título	1,263 (24%)
Proprietário sem título	3,095 (58%)
Inquilino	574 (11%)
Outro	419 (7.8%)
Parede em materiais definitivos	1,619 (30%)
Tecto em materiais definitivos	4,586 (86%)
Solo em materiais definitivos	2,737 (51%)

[1] n (%)

Characteristic	N = 5,351[1]
Logement du chef de ménage	
Proprietário com título	1,263 (24%)
Proprietário sem título	3,095 (58%)
Inquilino	574 (11%)
Outro	419 (7.8%)
Type de mur du logement	1,619 (30%)
Type de toit du logement	4,586 (86%)
Type de sol du logement	2,737 (51%)

[1] n (%)

```
library(labelled)
ehcvm_menage %>% labelled::to_factor() %>%select(logem,mur,toit,sol) %>% tbl_summary( l
 toit ~ "Type de toit du logement",
 mur~ "Type de mur du logement",
 sol~"Type de sol du logement"))
```

Le titre du tableau également n'est pas trop explicite, on va utiliser **modify_header()** pour l'adapter

```
library(labelled)
ehcvm_menage %>% labelled::to_factor() %>%select(logem,mur,toit,sol) %>% tbl_summary( l
 toit ~ "Type de toit du logement",
 mur~ "Type de mur du logement",
 sol~"Type de sol du logement")) %>% modify_header(label = "Caractérisques de l'habita
```

Pour les variables numériques, on peut également choisir ce qu'on veut calculer à travers la commande **statistic()** dans tbl_summary. Dans ce qui suit, on choisira de calculer la moyenne et l'écart des nouvelles variables *superf*, *grosrum* et *petitrum* intégrées.

```
library(labelled)
ehcvm_menage %>% labelled::to_factor() %>%select(logem,mur,toit,sol,superf,grosrum, pet
 grosrum ~ "Nombre de gros ruminants",
 petitrum~ "Nombre de petits ruminants",
```

Caractérisques de l'habitat du ménage	**N = 5,351**[1]
Logement du chef de ménage	
Proprietário com título	1,263 (24%)
Proprietário sem título	3,095 (58%)
Inquilino	574 (11%)
Outro	419 (7.8%)
Type de mur du logement	1,619 (30%)
Type de toit du logement	4,586 (86%)
Type de sol du logement	2,737 (51%)

[1] n (%)

Agriculture, Elevage et logement	**N = 5,351**[1]
Logement du chef de ménage	
Proprietário com título	1,263 (24%)
Proprietário sem título	3,095 (58%)
Inquilino	574 (11%)
Outro	419 (7.8%)
Type de mur du logement	1,619 (30%)
Type de toit du logement	4,586 (86%)
Type de sol du logement	2,737 (51%)
Superficie agricole	6,698,899.56 (400,815,757.59)
Unknown	1,771
Nombre de gros ruminants	1.8 (9.0)
Nombre de petits ruminants	2.4 (6.3)

[1] n (%); Mean (SD)

```
logem ~ "Logement du chef de ménage",
toit ~ "Type de toit du logement",
mur~ "Type de mur du logement",
sol~"Type de sol du logement"
),
statistic = list(superf ~ "{mean} ({sd})",##pour avoir la moyenne et l'écart-type
                 grosrum ~ "{mean} ({sd})",
                 petitrum ~ "{mean} ({sd})"
                 )
) %>% modify_header(label = "Agriculture, Elevage et logement")
```

Pour choisir le nombre de chiffres après la virgule des différents indicateurs, on peut utiliser la commande **digits = everything() ~c(0,0,0,0,…)** dans tbl_summary où everything() sélectionne toutes les colonnes du tableau et c(0,0,0,0,…) définit un format spécifique pour chaque colonne, en indiquant 0 décimale pour chacun

```
library(labelled)
ehcvm_menage %>% labelled::to_factor() %>%select(logem,mur,toit,sol,superf,grosrum, pet
 grosrum ~ "Nombre de gros ruminants",
```

Agriculture, Elevage et logement	**N = 5,351**[1]
Logement du chef de ménage	
Proprietário com título	1,263 (24%)
Proprietário sem título	3,095 (58%)
Inquilino	574 (11%)
Outro	419 (8%)
Type de mur du logement	1,619 (30%)
Type de toit du logement	4,586 (86%)
Type de sol du logement	2,737 (51%)
Superficie agricole	6,698,900 (400,815,758)
Unknown	1,771
Nombre de gros ruminants	2 (9)
Nombre de petits ruminants	2 (6)

[1]n (%); Mean (SD)

```
 petitrum~ "Nombre de petits ruminants",
 logem ~ "Logement du chef de ménage",
 toit ~ "Type de toit du logement",
 mur~ "Type de mur du logement",
 sol~"Type de sol du logement"
 ),
 statistic = list(superf ~ "{mean} ({sd})",##pour avoir la moyenne et l'écart-type
                  grosrum ~ "{mean} ({sd})",
                  petitrum ~ "{mean} ({sd})"
                  ),
 digits = everything() ~c(0,0,0,0)
 ) %>% modify_header(label = "Agriculture, Elevage et logement")
```

Pour les valeurs manquantes, on peut afficher leur nombre pour chaque variable avec missing = “always” et changer leur appelation par *Valeurs manquantes* ou *NA* avec la commande **missing_text()**

```
library(labelled)
ehcvm_menage %>% labelled::to_factor() %>%select(logem,mur,toit,sol,superf,grosrum, pet
 grosrum ~ "Nombre de gros ruminants",
 petitrum~ "Nombre de petits ruminants",
 logem ~ "Logement du chef de ménage",
 toit ~ "Type de toit du logement",
 mur~ "Type de mur du logement",
 sol~"Type de sol du logement"
 ),
 statistic = list(superf ~ "{mean} ({sd})",##pour avoir la moyenne et l'écart-type
                  grosrum ~ "{mean} ({sd})",
                  petitrum ~ "{mean} ({sd})"
                  ),
 digits = everything() ~c(0,0,0,0),
 missing = "always",  ##Pour afficher les missings
 missing_text= "Valeurs manquantes"
 ) %>% modify_header(label = "Agriculture, Elevage et logement")
```

Agriculture, Elevage et logement	**N = 5,351**[1]
Logement du chef de ménage	
Proprietário com título	1,263 (24%)
Proprietário sem título	3,095 (58%)
Inquilino	574 (11%)
Outro	419 (8%)
Valeurs manquantes	0
Type de mur du logement	1,619 (30%)
Valeurs manquantes	0
Type de toit du logement	4,586 (86%)
Valeurs manquantes	0
Type de sol du logement	2,737 (51%)
Valeurs manquantes	0
Superficie agricole	6,698,900 (400,815,758)
Valeurs manquantes	1,771
Nombre de gros ruminants	2 (9)
Valeurs manquantes	0
Nombre de petits ruminants	2 (6)
Valeurs manquantes	0

[1]n (%); Mean (SD)

Le tableau ainsi prêt, essayons d'analyser les statistiques descriptives des variables choisies dans la base ehcvm_ménage.

Analyse descriptive de quelques variables pour les deux bases

Le tableau présente des données sur 5 351 ménages concernant leur logement et leurs activités agricoles.

La majorité des chefs de ménage sont propriétaires (82% (58+24)), principalement sans titre (58%), tandis que 11% sont locataires. Les caractéristiques du logement montrent que 30% ont un type spécifique de mur, 86% un type particulier de toit et 51% un certain type de sol, sans valeurs manquantes.

Concernant l'agriculture, la superficie agricole a une moyenne élevée de 6 698 900 avec un écart-type très important (400 815 758), indiquant une grande dispersion, et 1 771 valeurs manquantes. L'élevage révèle une moyenne de 2 gros ruminants (écart-type 9) et 2 petits ruminants (écart-type 6), sans valeurs manquantes.

Faisons de même pour la base ehcvm_welfare :

```
library(haven)
ehcvm_welfare %>% labelled::to_factor() %>% select(hgender,hage,hmstat,heduc,hdiploma)
  label = list(hgender ~ "Genre du chef de ménage",
 hage ~ "Âge du chef de ménage",
 hmstat~ "Situation matrimoniale du chef de ménage",
 heduc ~ "Niveau d'éducation du chef de ménage",
 hdiploma ~ "Diplôme du chef de ménage"
```

```
),
statistic = list(hage ~ "{mean} ({sd})",
                 hgender~ "{n}/{N} ({p}%)"##pour avoir la moyenne et l'écart-type
                 ),
digits = everything() ~c(0,0,0,0),
missing = "always",  ##Pour afficher les missings
missing_text= "Valeurs manquantes"
) %>% modify_header(label = "Caractéristiques du chef de ménage")
```

Le tableau ci-dessus présente des données sur les caractéristiques des chefs de ménage. La majorité sont des hommes (78%), avec un âge moyen de 50 ans (écart-type 14). Concernant la situation matrimoniale, 51% sont mariés monogames, 20% polygames, 12% célibataires, et 13% veufs. En termes d'éducation, 47% n'ont aucun niveau scolaire, tandis que 18% ont un premier cycle primaire et 12% un deuxième cycle. Concernant les diplômes, 56% n'en possèdent aucun, 16% ont un diplôme de premier cycle primaire et 10% de deuxième cycle, alors que les niveaux supérieurs (licence, master, doctorat) restent marginaux. Il n'y a aucune valeur manquante dans les données.

Caractéristiques du chef de ménage	**N = 5,351**[1]
Genre du chef de ménage	
Masculino	4,161/5,351 (78%)
Feminino	1,190/5,351 (22%)
Valeurs manquantes	0
Âge du chef de ménage	50 (14)
Valeurs manquantes	0
Situation matrimoniale du chef de ménage	
Solteiro (a)	628 (12%)
Casado(a) monogâmico(a)	2,703 (51%)
Casado(a) poligâmico(a)	1,064 (20%)
União de facto	62 (1%)
Viúvo(a)	706 (13%)
Divorciado(a)	44 (1%)
Separado(a)	144 (3%)
Valeurs manquantes	0
Niveau d'éducation du chef de ménage	
Nenhum	2,497 (47%)
Pre-escolar	70 (1%)
primario 1 ciclo	948 (18%)
Primario 2 ciclo	668 (12%)
primario 3 ciclo.gl 1	424 (8%)
Ensino. tec.Prof 1	49 (1%)
Second. gl 2	318 (6%)
Ensino Medio	223 (4%)
Superior	154 (3%)
Valeurs manquantes	0
Diplôme du chef de ménage	
Nenhum	3,013 (56%)
EP 1 (Ensino Primário 1º ciclo)	840 (16%)
EP 2 (Ensino Primário 2º ciclo)	521 (10%)
EP 3 (Ensino Primário 3º ciclo)	332 (6%)
ETP (Ensino Técnico/Profissional)	46 (1%)
ES (Ensino Secundário)	256 (5%)
EM (Ensino Médio)	207 (4%)
Licenciado	119 (2%)
Pós graduação/Especialização	2 (0%)
Mestrado	6 (0%)
Doutorado	9 (0%)
Valeurs manquantes	0

[1] n/N (%); Mean (SD); n (%)