Projet- logiciel statistique R: Le package gtsumary

"Groupe 12: ADAM Alassane, CISSE Pape Abdourahmane, NGOM Fallou, YAMIN Youdan"

Dernière mise à jour: 26-mai-2024

République du Sénégal

Un Peuple-Un But-Une Foi

Ministère de l'Economie, du Plan et de la Coopération

Agence nationale de la Statistique et de la Démographie

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

Projet statistiques sous R

Package Gtsummary

Rédigé par :

ADAM Alassane, CISSE Pape Abdourahmane, NGOM Fallou, KONLAMBIGUE Youdan-yamin

Elèves ingénieurs statisticiens économistes en troisième année de classe préparatoire

Professeur :

M. HEMA

Année scolaire 2023-2024

Contents

Note

La bibliothèque GTsummary est une puissante extension de R pour la création de tableaux de synthèse des résultats d'analyses statistiques. Son objectif principal est de simplifier et d'améliorer la création de rapports statistiques en fournissant des outils conviviaux pour résumer et présenter les résultats de manière claire et concise.
Pour utiliser gtsumary: Importer les données dans R, Utiliser les fonctions gtsummary pour résumer les resultats des analyses statistiques, personnaliser les tableaux générés selon vos besoins en utilisant les options de personnalisations disponibles, Intégrer les tableaux dans vos rapports ou présentations pour une communication claire et efficace des résultats.

1 I Préliminaires

1.1 I-1 Utilité

Le package `gtsummary` en R est un outil puissant conçu pour la génération de tableaux de synthèse de données statistiques de manière élégante, personnalisable, reproductible et directement publiables. Il intègre :

- **Tableaux descriptifs** : univariés, bivariés et de regression
- **Résumés des données descriptives** : moyennes, médianes, écarts-types, fréquence ...
- **Test statistiques** : t-test, khi-2...
- **Personnalisation avancé** : apparence, titres, légendes, annotations ...
- **Reproductibilité** : Automatisation

1.2 I-2 Installation et types de variables

1.2.1 I-2-1 Installation

- Installer depuis le CRAN :

```
install.packages("gtsummary")
```

- Installer la version de dévellopement depuis github

```
remotes::install_github("ddsjoberg/gtsummary")
```

1.2.2 I-2-2 Types de variables

Il y'a trois types de variables dans gtsummary. Par défaut, gtsummary considère qu'une variable est :

- **Dichotomique** : s'il s'agit d'un vecteur logique (TRUE/FALSE), d'une variable textuelle codée yes/no ou d'une variable numérique codée 0/1.
- **Catégorielle** : S'il s'agit d'un facteur, d'une variable textuelle ou d'une variable numérique ayant moins de 10 valeurs différentes.
- **Continue** : Dans les autres cas de variables numériques.

1.3 I-3 Description de la base de données

- **Source** : EHCVM 2018
- **Taille** : 66120 ménages enquêtés et 11 variables
- **Description** : Les variables renseignent sur la localisation des ménages, les carctérisques sociodémographiques du chef de ménage, la consommation annulle et les indicateurs de pauvreté.

Variable	Lalel	class
region	Région de résidence	factor
milieu	Milieu de résidence	factor
hgender	genre chef ménage	factor
hbranch	branche d'activité chef de ménage	factor
catégorie_hhsize	taille du ménage	factor
categorie_age	age du chef de ménage	factor
dali	Consommation annuelle	numeric
p0	statut pauvreté	factor
P11	profondeur de la pauvreté	numeric
P22	sévérité de la pauvreté	numeric

2 II Tableau descriptif univarié

2.1 II-1 Thémes du tableaux

gtsummary fournit plusieurs fonctions préfixées **theme_gtsummary()** permettent de modifier l'affichage par défaut des tableaux. Parmis les Exemples de fonctions nous avons: La fonction **theme_gtsummary_language**(Permets de modifier la langue utlisés dans le tableaux), La fonction **theme_gtsummary_journal**(Pour définir un thème prédéfini), La fonction **reset_gtsummary_theme**(Pour effacer tous les thémes précédemment définis)

```
#Ramener le format du tableau en français
theme_gtsummary_language(language = "fr", decimal.mark = ",",
                         big.mark = " ")
```

2.2 II-2 La fonction tbl_summary

La fonction au coeur du package **gtsummary** se nomme **tbl_summary()**.Elle produit un tableau qui s'affiche dans l'onglet "Viewer". On lui passe en entrée un tableaux de données (data.frame) et par défaut toutes les variables sont résumé (base %>%tbl_summary())

La fonction *tbl_summary* permet d'obtenir la statistique descriptive ou tri à plat des variables, c'est-à-dire obtenir la moyenne, l'écart-type, intervalle interquartile, etc.Elle prend au minimum une base de données. Dans ce cas, elle affiche des statistiques descriptives pour chaque variable.

2.2.1 II-2-1 Sélection des variables (include)

La paramètre include permets de spécifier les variables à inclure dans le tableau (et leur ordre). On peut lui passer un vecteur de noms de variables, ou bien utiliser des sélecteurs tidyselect (utiliser c() si plusieurs sélecteurs).

Par exemple, le tableau suivant nous donne des statistiques descriptives sur les variables **milieu**, **Niveau d'éducation du chef du ménage**, la variable **genre du chef de ménage**

```
# Appliquer la fonction tbl_summary() à l'ensemble de données 'basev'
basev %>%
  tbl_summary(include = c("milieu","hgender","heduc"),

  value = list( milieu ~ "Rural", hgender ~ "Masculin"),
  label = list(milieu ~ "Milieu Rural",hgender ~ "Genre Masculin"))
```

Caractéristique	**N = 66 120**
Milieu Rural	32 798 (50%)
Genre Masculin	51 644 (78%)

Caractéristique	**N = 66 120**
Education du CM	
Aucun	47 811 (72%)
Maternelle	6 (<0,1%)
Primaire	9 019 (14%)
Second. gl 1	4 178 (6,3%)
Second. tech. 1	405 (0,6%)
Second. gl 2	2 002 (3,0%)
Second. tech. 2	187 (0,3%)
Postsecondaire	425 (0,6%)
Superieur	2 087 (3,2%)

Il important de souligner que les statistiques envoyées par défaut dépendent du type de la variable. Ainsi, pour les variables du type numérique, nous avons un format du type : mediane (intervalle interquartile). Toutefois, il est possible de paramétrer ces arguments par défauts.

2.2.2 II-2-2 Etiquettes des variables

Pour modifier l'étiquettes associé à une certaine variable,on peut utliser l'option *label* de *tbl_summary* par exemple:

```
# Résumé tabulaire des statistiques descriptives pour les variables

'milieu'#, hgender avec une étiquette personnalisée
```

```
## [1] "milieu"
```

```
basev %>%
  tbl_summary(include = c("milieu","hgender", "hbranch"),label =
                list(milieu~"milieu de naissance du CM", hgender~"genre CM")  )
```

Caractéristique	**N = 66 120**
milieu de naissance du CM	
Urbain	33 322 (50%)
Rural	32 798 (50%)
genre CM	
Masculin	51 644 (78%)
Féminin	14 476 (22%)
Branche activite du CM	
Agriculture	15 199 (31%)
Elevage/peche	3 371 (6,8%)
Indust. extr.	494 (1,0%)
Autr. indust.	4 187 (8,5%)
BTP	2 510 (5,1%)
Commerce	9 441 (19%)
Restaurant/Hotel	431 (0,9%)
Trans./Comm.	2 250 (4,6%)
Education/Sante	2 894 (5,9%)
Services perso.	6 561 (13%)
Aut. services	2 004 (4,1%)
Manquant	16 778

Il est également possible d'utliser la syntaxe *tidyselect* et les selecteurs de *tidyselect* comme *everything*,*starts_with*,*contains* ou*all_of*.

```
# Résumé tabulaire des statistiques descriptives pour toutes les variables de l'ensemble de données
'base' #avec une étiquette commune
```

```
## [1] "base"
```

```
basev %>%
  tbl_summary(include = c("milieu","hgender","heduc"),label=everything()~"Etiquette")
```

Caractéristique	**N = 66 120**
Etiquette	
Urbain	33 322 (50%)
Rural	32 798 (50%)
Etiquette	
Masculin	51 644 (78%)
Féminin	14 476 (22%)
Etiquette	
Aucun	47 811 (72%)
Maternelle	6 (<0,1%)
Primaire	9 019 (14%)
Second. gl 1	4 178 (6,3%)
Second. tech. 1	405 (0,6%)
Second. gl 2	2 002 (3,0%)
Second. tech. 2	187 (0,3%)
Postsecondaire	425 (0,6%)
Superieur	2 087 (3,2%)

2.3 II-3 Statistiques à afficher

on peut définir une liste dans laquelle on indique des formules spécifient les types de statistiques descriptives à afficher pour les variables ,comme suit:

```
# Générer un résumé tabulaire des statistiques
#descriptives pour les variables 'P1' et 'P22' de
#l'ensemble de données 'basev'
# Spécifier les statistiques à afficher pour les variables P11, P22
basev %>%
  tbl_summary(
    include = c(P11,P22,dali),  # Sélectionner la variable 'region'
    statistic = all_continuous() ~ "Moy.:{mean}[min-max:{min}-{max}]",
    #Spécifier les statistiques à afficher
 label = list(P11~"pronfondeur de la pauvreté",P22~"Sévérité de la
              pauvreté", dali~"Consommation annuelle") )
```

Caractéristique	**N = 66 120**
pronfondeur de la pauvreté	Moy.:0,12[min-max:0,00-0,81]
Sévérité de la pauvreté	Moy.:0,05[min-max:0,00-0,65]
Consommation annuelle	Moy.:2 513 834[min-max:113 187-31 295 272]

Il est possible d'afficher des statistiques différentes pour chaque variable.

```r
# Générer un résumé tabulaire des statistiques descriptives pour les
#variables 'region' et 'milieu' de l'ensemble de données 'base'
# Trier les modalités des variables catégorielles par fréquence,
#de la plus
#fréquente à la moins fréquente
basev %>%
  tbl_summary(
    include = c(region, milieu),
    # Sélectionner les variables region et milieu
    sort = all_categorical() ~ "frequency"
    # Trier les modalités par fréquence pour les variables catégorielles
  )
```

Caractéristique	**N = 66 120**
Region residence	
DAKAR	7 116 (11%)
DIOURBEL	5 473 (8,3%)
THIES	5 457 (8,3%)
KAOLACK	5 374 (8,1%)
SAINT-LOUIS	4 998 (7,6%)
LOUGA	4 729 (7,2%)
TAMBACOUNDA	4 397 (6,7%)
KAFFRINE	4 367 (6,6%)
SEDHIOU	4 329 (6,5%)
MATAM	4 197 (6,3%)
KOLDA	4 085 (6,2%)
FATICK	4 084 (6,2%)
KEDOUGOU	3 953 (6,0%)
ZIGUINCHOR	3 561 (5,4%)
Milieu residence	
Urbain	33 322 (50%)
Rural	32 798 (50%)

La fonction **add_n()** est utilisée pour ajouter une colonne au résumé tabulaire qui indique le nombre d'observations non manquantes pour chaque variable par défaut.

```r
# Générer un résumé tabulaire des statistiques descriptives pour les
#variables 'region' et 'milieu' de l'ensemble de données 'base'
# Ajouter une colonne avec le nombre d'observations non manquantes par défaut
basev %>%
  tbl_summary(include = c(region, milieu)) %>%
  add_n()
```

Caractéristique	**N**	**N = 66 120**
Region residence	66 120	
DAKAR		7 116 (11%)
ZIGUINCHOR		3 561 (5,4%)
DIOURBEL		5 473 (8,3%)
SAINT-LOUIS		4 998 (7,6%)
TAMBACOUNDA		4 397 (6,7%)
KAOLACK		5 374 (8,1%)
THIES		5 457 (8,3%)

Caractéristique	N	N = 66 120
LOUGA		4 729 (7,2%)
FATICK		4 084 (6,2%)
KOLDA		4 085 (6,2%)
MATAM		4 197 (6,3%)
KAFFRINE		4 367 (6,6%)
KEDOUGOU		3 953 (6,0%)
SEDHIOU		4 329 (6,5%)
Milieu residence	66 120	
Urbain		33 322 (50%)
Rural		32 798 (50%)

```
# Ajouter une colonne avec le nombre d'observations non
#manquantes par défaut
```

2.4 II-4 Intervalle de confiance(add_ci)

l'argument **add_ci()** est utilisée pour ajouter les intervalles de confiance au résumé tabulaire.

```
# Générer un résumé tabulaire des statistiques descriptives pour les
#variables 'region' et 'milieu' de l'ensemble de données 'basev'
# Ajouter une colonne avec le nombre d'observations non manquantes par défaut
basev %>%
  tbl_summary(include = c(region, milieu)) %>%
  add_ci() # Ajouter les intervalles de confiance
```

Caractéristique	N = 66 120	95% CI
Region residence		
DAKAR	7 116 (11%)	11%, 11%
ZIGUINCHOR	3 561 (5,4%)	5,2%, 5,6%
DIOURBEL	5 473 (8,3%)	8,1%, 8,5%
SAINT-LOUIS	4 998 (7,6%)	7,4%, 7,8%
TAMBACOUNDA	4 397 (6,7%)	6,5%, 6,8%
KAOLACK	5 374 (8,1%)	7,9%, 8,3%
THIES	5 457 (8,3%)	8,0%, 8,5%
LOUGA	4 729 (7,2%)	7,0%, 7,4%
FATICK	4 084 (6,2%)	6,0%, 6,4%
KOLDA	4 085 (6,2%)	6,0%, 6,4%
MATAM	4 197 (6,3%)	6,2%, 6,5%
KAFFRINE	4 367 (6,6%)	6,4%, 6,8%
KEDOUGOU	3 953 (6,0%)	5,8%, 6,2%
SEDHIOU	4 329 (6,5%)	6,4%, 6,7%
Milieu residence		
Urbain	33 322 (50%)	50%, 51%
Rural	32 798 (50%)	49%, 50%

2.5 II-5 Données manquantes

Le package gtsummary offre plusieurs paramétres pour manipuler les données manquantes ,presenter ci-dessous:

```
basev %>%
  tbl_summary(
```

```
    # Inclure les colonnes "milieu" et "region" dans le résumé
    include = c("milieu", "region"),
    # Indiquer qu'il faut toujours afficher le nombre
    #d'observations manquantes
    missing = "always",
    # Personnaliser le texte affiché pour les observations manquantes
    missing_text = "Nbre observations manquantes"
  )
```

Caractéristique	**N = 66 120**
Milieu residence	
Urbain	33 322 (50%)
Rural	32 798 (50%)
Nbre observations manquantes	0
Region residence	
DAKAR	7 116 (11%)
ZIGUINCHOR	3 561 (5,4%)
DIOURBEL	5 473 (8,3%)
SAINT-LOUIS	4 998 (7,6%)
TAMBACOUNDA	4 397 (6,7%)
KAOLACK	5 374 (8,1%)
THIES	5 457 (8,3%)
LOUGA	4 729 (7,2%)
FATICK	4 084 (6,2%)
KOLDA	4 085 (6,2%)
MATAM	4 197 (6,3%)
KAFFRINE	4 367 (6,6%)
KEDOUGOU	3 953 (6,0%)
SEDHIOU	4 329 (6,5%)
Nbre observations manquantes	0

2.6 II-6 Statistique personnalisée avec tbl_continous et tbl_custom_summary

2.6.1 II-6-1 tbl_contunous()

La fonction tbl_continuous permets de résumer une variable continue en fonction de deux ou plusieurs variables catégorielles.

Par exemple, pour afficher la consommation moyenne moyen de plusieurs sous-groupes :

```
# Afficher pour chaque milieu la moyenne dali
basev%>%
  tbl_continuous(
    variable = dali,
    statistic = ~ "{mean}",
    include = milieu
  )
```

Caractéristique	**N = 66 120**
Milieu residence	
Urbain	2 854 149
Rural	2 168 082

2.6.2 II-6-2 tbl_custom_summary()

La fonction tbl_custom_summary permets encore plus de personnalisation que tbl_continuous .

On doit fournir via stat_fns une fonction personnalisée qui va recevoir un sous tableau de données , contenant toutes les variables du fichier, et qui renverra des statistiques personnalisées que l'on affichera avec statistic . La fonction peut-être différente pour chaque variable. Il est également possible d'utiliser quelques fonctions dédiées fournies directement par gtsummary.

```
# afficher pour chaque milieu, le genre, la taille
#du ménage la proportion de pauvreté
basev%>%
  tbl_custom_summary(
  include = c(milieu, hgender, categorie_hhsize),
  stat_fns = ~proportion_summary(variable="p0",value="pauvre"),
  statistic = ~"{prop}"
)
```

Caractéristique	**N = 66 120**
Milieu residence	
Urbain	0,29
Rural	0,55
Genre du chef de ménage	
Masculin	0,47
Féminin	0,26
Taille du ménage	
Moins de 4 persones	0,06
5 à 9 persones	0,29
10 à 14 persones	0,43
15 à 19 persones	0,52
20 persones et plus	0,70

2.7 II-7 Application

```
#Application 1
basev %>%
  tbl_summary(include = c(milieu,hgender,dali),
              missing = "always",
              missing_text = "Nbre observations manquantes",
              sort = all_categorical() ~ "frequency",
              value = list( milieu ~ "Rural", hgender ~ "Masculin"),
              label = list(milieu ~ "Milieu Rural",
                           hgender ~ "Genre Masculin",
                           dali~"Consommation annuelle"),
              statistic = all_continuous() ~
                "Moy.:{mean}[min-max:{min}-{max}]",) %>%
              add_n()  %>%
              add_ci()
```

Caractéristique	**N**	**N = 66 120**	**95% CI**
Milieu Rural	66 120	32 798 (50%)	49%, 50%
Nbre observations manquantes		0	

Caractéristique	**N**	**N = 66 120**	**95% CI**
Genre Masculin	66 120	51 644 (78%)	78%, 78%
Nbre observations manquantes		0	
Consommation annuelle	66 120	Moy.:2 513 834[min-max:113 187-31 295 272]	2 500 157, 2 527 511
Nbre observations manquantes		0	

```r
#Application 2
## Créer un dataframe pour les années passées
my_data <- data.frame(
  milieu = c("Urbain", "Rural"),
  taux = c(0.22, 0.58)
)

## Traduire le dataframe en tableau gt
pauvreté2011<-my_data%>%
  mutate(milieu=factor(milieu, levels=c("Urbain","Rural")))%>%
  tbl_custom_summary(
  include = c(milieu), label = milieu~ "Milieu residence",
  stat_fns = ~continuous_summary("taux"),
  statistic = ~"{mean}",
  digits = ~ list(
      function(x) {
        style_percent(x, digits = 1)
      },
      0, 0, style_percent, style_percent
    ),
  overall_row = TRUE,  ##
  overall_row_last = TRUE
)%>%
  modify_header(stat_0~"**%**")%>%
  modify_footnote(everything()~NA)

## Le tableau de la proportion de pauvres
pauvreté2018<-basev%>%
  tbl_custom_summary(
  include = c(milieu),
  stat_fns = ~proportion_summary(variable="p0",value="pauvre"),
  statistic = ~"{prop}",
  digits = ~ list(
      function(x) {
        style_percent(x, digits = 1)
      },
      0, 0, style_percent, style_percent
    ),
  overall_row = TRUE,
  overall_row_last = TRUE
)%>%
  modify_header(stat_0~"**%**")%>%
  modify_footnote(everything()~NA)
```

```
## Merger les deux tableaux
tbl_merge(list(pauvreté2011,pauvreté2018),tab_spanner =
            c("Taux de Pauvreté 2011",
              "Taux de Pauvreté 2018/2019"))%>%
  as_gt()%>%
  gt::tab_header(
    title=gt::md("**Taux de pauvreté selon le milieu**"))%>%
  gt::tab_source_note("EHCVM, calculs de l'auteur")
```

Taux de pauvreté selon le milieu

	Taux de Pauvreté 2011	Taux de Pauvreté 2018/2019
Caractéristique	%	%
Milieu residence		
Urbain	22,0	29,4
Rural	58,0	55,2
Total	40,0	42,2

EHCVM, calculs de l'auteur

```
# Application 3
## Pauvreté selon le niveau d'éducation
### Tableau de l'incidence de pauvreté
incidence<-basev%>%
  tbl_custom_summary(
  include = c(heduc),
  stat_fns = ~proportion_summary(variable="p0",value="pauvre"),
  statistic = ~"{prop}",
  digits = ~ list(
      function(x) {
        style_percent(x, digits = 1)
      },
      0, 0, style_percent, style_percent
    )
)%>%
  bold_labels()%>%
  italicize_levels()%>%
  modify_header(stat_0~"**%**")%>%
  modify_footnote(everything()~NA)
### Le tableau de la profonduer de pauvreté
profondeur<-basev%>%
  tbl_custom_summary(
  include = c(heduc),
  stat_fns = ~continuous_summary("P11"),
  statistic = ~"{mean}",
  digits = ~ list(
      function(x) {
        style_percent(x, digits = 1)
      },
      0, 0, style_percent, style_percent
    )
)%>%
  bold_labels()%>%
```

```
  italicize_levels()%>%
  modify_header(stat_0~"**%**")%>%
  modify_footnote(everything()~NA)
### Le tableau de la sévérité de pauvreté
severite<-basev%>%
  tbl_custom_summary(
  include = c(heduc),
  stat_fns = ~continuous_summary("P22"),
  statistic = ~"{mean}",
  digits = ~ list(
      function(x) {
        style_percent(x, digits = 1)
      },
      0, 0, style_percent, style_percent
    )
)%>%
  bold_labels()%>%
  italicize_levels()%>%
  modify_header(stat_0~"**%**")%>%
  modify_footnote(everything()~NA)
###Merger les trois tableaux
tbl_merge(list(incidence,profondeur,severite),tab_spanner =
            c("Incidence","profondeur","sévérité"))%>%
  as_gt()%>%
  gt::tab_header(
    title=gt::md("**Indicateur de pauvreté selon la niveau
                  d'éducation**"))%>%
  gt::tab_source_note("EHCVM, calculs de l'auteur")
```

Indicateur de pauvreté selon la niveau d'éducation

	Incidence	profondeur	sévérité
Caractéristique	%	%	%
Education du CM			
Aucun	49,0	14,4	5,83
Maternelle	0	0	0
Primaire	31,1	7,59	2,66
Second. gl 1	24,8	6,43	2,34
Second. tech. 1	35,3	9,18	3,16
Second. gl 2	13,1	3,24	1,03
Second. tech. 2	24,1	1,92	0,15
Postsecondaire	0	0	0
Superieur	7,19	0,85	0,27

EHCVM, calculs de l'auteur

```
# Application 4
## Pauvreté selon la région
### Tableau de l'incidence de pauvreté
incidence<-basev%>%
  tbl_custom_summary(
  include = c(region),
  stat_fns = ~proportion_summary(variable="p0",value="pauvre"),
```

```
  statistic = ~"{prop}",
  digits = ~ list(
      function(x) {
        style_percent(x, digits = 1)
        # Metrre les proportion en format %
      },
      0, 0, style_percent, style_percent
    )
)%>%
  bold_labels()%>%
  italicize_levels()%>%
  modify_header(stat_0~"**%**")%>%
  modify_footnote(everything()~NA)
### Le tableau de la profonduer de pauvreté
profondeur<-basev%>%
  tbl_custom_summary(
  include = c(region),
  stat_fns = ~continuous_summary("P11"),
  statistic = ~"{mean}",
  digits = ~ list(
      function(x) {
        style_percent(x, digits = 1)
        # Metrre les proportion en format %
      },
      0, 0, style_percent, style_percent
    )
)%>%
  bold_labels()%>%
  italicize_levels()%>%
  modify_header(stat_0~"**%**")%>%
  modify_footnote(everything()~NA)
### Le tableau de la sévérité de pauvreté
severite<-basev%>%
  tbl_custom_summary(
  include = c(region),
  stat_fns = ~continuous_summary("P22"),
  statistic = ~"{mean}",
  digits = ~ list(
      function(x) {
        style_percent(x, digits = 1)
        # Metrre les proportion en format % à un
        #chiffre aprés la virgule
      },
      0, 0, style_percent, style_percent
    )
)%>%
  bold_labels()%>%
  italicize_levels()%>%
  modify_header(stat_0~"**%**")%>%
  modify_footnote(everything()~NA)
###Merger les trois tableaux
tbl_merge(list(incidence,profondeur,severite),tab_spanner =
            c("Incidence","profondeur","sévérité"))%>%
```

```
  as_gt()%>%
  gt::tab_header(
    title=gt::md("**Indicateur de pauvreté selon le niveau
                 d'éducation**"))%>%
  gt::tab_source_note("EHCVM, calculs de l'auteur")
```

Indicateur de pauvreté selon le niveau d'éducation

Caractéristique	Incidence %	profondeur %	sévérité %
Region residence			
DAKAR	9,74	1,43	0,35
ZIGUINCHOR	47,9	14,3	5,90
DIOURBEL	43,9	10,3	3,41
SAINT-LOUIS	39,4	11,1	4,32
TAMBACOUNDA	58,7	18,0	7,43
KAOLACK	35,9	10,3	4,15
THIES	34,9	7,96	2,59
LOUGA	36,3	9,40	3,37
FATICK	42,0	11,4	4,08
KOLDA	54,8	15,7	6,12
MATAM	48,9	15,1	6,46
KAFFRINE	48,2	15,2	6,61
KEDOUGOU	55,8	19,0	8,82
SEDHIOU	61,7	19,6	8,13

EHCVM, calculs de l'auteur

3 III Tableaux croisés

Il s'agit dans cette partie de savoir comment ventiler les fréquences de deux variables catégorielles dans un tableau, comment faire sortir les fréquences et éventuellement utliser quelques fonctions du package **gtsummary** telles que les thèmes...

3.1 III-1 Tableaux croisés avec tbl_summary et tbl_custom_summary

Nous allons à présent utiliser les fonctions tbl_summary et tbl_custom_summary combinées avec **by**. Il s'agit dans cette partie d'analyser la pauvreté en fonction du genre du chef du ménage, de la région, du milieu de résidence ... Il est imporatant de savoir que le regroupement se fait par une variable catégorielle sinon, le regroupement n'aura pas de sens.

```
basev %>%
  tbl_summary(include = c(P11, P22),
    by = hgender,
    label = list(P11 ~ "Profondeur de pauvreté",
                 P22 ~ "Sévérité de la pauvreté"),
    statistic = list(c(P11, P22)~ "{mean}")
)%>%  add_difference() %>%
  as_gt()%>%
  gt::tab_header(
    title=gt::md("**Profondeur et sévérité de la
                 pauvreté selon le sexe CM**"))%>%
  gt::tab_source_note("EHCVM, calculs de l'auteur")
```

Profondeur et sévérité de la pauvreté selon le sexe CM

Caractéristique	Masculin, N = 51 644[1]	Féminin, N = 14 476[1]	Difference[2]	95% IC[2,3]	p-valeur[2]
Profondeur de pauvreté	0,14	0,06	0,07	0,07 – 0,08	<0,001
Sévérité de la pauvreté	0,05	0,02	0,03	0,03 – 0,03	<0,001

[1]Moyenne
[2]test de Student
[3]IC = intervalle de confiance
EHCVM, calculs de l'auteur

Analyse de la pauvreté selon l'âge et le genre du chef de ménage

```
basev %>%
  tbl_custom_summary(
    include = "categorie_age",
    label = categorie_age~ "Classe d'âge",
    by = "hgender",
    stat_fns = ~ proportion_summary("p0", "pauvre"),
    statistic = ~"{prop}% ",
    digits = ~ list(
      function(x) {
        style_percent(x, digits = 1)
      },
      0, 0, style_percent, style_percent
    ),
    overall_row = TRUE,
    overall_row_last = TRUE
  ) %>%
  bold_labels() %>%
  modify_footnote(
    update = all_stat_cols() ~ ""
  )%>%
  as_gt()%>%
  gt::tab_header(
    title=gt::md("**Taux de pauvreté selon l'âge et le sexe**"))%>%
  gt::tab_source_note("EHCVM, calculs de l'auteur")
```

Taux de pauvreté selon l'âge et le sexe

Caractéristique	Masculin, N = 51 644[1]	Féminin, N = 14 476[1]
Classe d'âge		
Moins de 24 ans	50,0%	28,5%
25 à 39 ans	41,9%	21,6%
40 à 49 ans	41,2%	19,3%
50 à 59 ans	40,5%	22,5%
60 ans et plus	40,3%	19,9%
Manquant	4	3
Total	46,8%	25,8%

[1]
EHCVM, calculs de l'auteur

3.2 III-2 Tableaux croisés avec tbl_cross

Milieu selon la catégorie d'âge.

```
basev %>% tbl_cross(row = categorie_age,
                    col = categorie_hhsize,
                    percent = "cell",
                    label = categorie_age ~ "Age CM" ) %>%

  as_gt()%>%
  gt::tab_header(
    title=gt::md("**Répartition des ménages
                  selon la taille et l'age CM**"))%>%
  gt::tab_source_note("EHCVM, calculs de l'auteur")
```

Répartition des ménages selon la taille et l'age CM

	Taille du ménage				
	Moins de 4 persones	5 à 9 persones	10 à 14 persones	15 à 19 persones	20 persones et plus
Age CM					
Moins de 24 ans	1 493 (2,3%)	13 471 (20%)	12 704 (19%)	7 235 (11%)	7 091 (11%)
25 à 39 ans	847 (1,3%)	3 605 (5,5%)	3 249 (4,9%)	1 801 (2,7%)	1 757 (2,7%)
40 à 49 ans	425 (0,6%)	1 863 (2,8%)	1 369 (2,1%)	721 (1,1%)	710 (1,1%)
50 à 59 ans	335 (0,5%)	1 352 (2,0%)	1 002 (1,5%)	488 (0,7%)	434 (0,7%)
60 ans et plus	376 (0,6%)	1 412 (2,1%)	1 213 (1,8%)	621 (0,9%)	539 (0,8%)
Unknown	0 (0%)	2 (<0,1%)	3 (<0,1%)	1 (<0,1%)	1 (<0,1%)
Total	3 476 (5,3%)	21 705 (33%)	19 540 (30%)	10 867 (16%)	10 532 (16%)

EHCVM, calculs de l'auteur

4 Iv Régression logistique binaire

La régression logistique binaire (également appelé modèle logit) est souvent utilisé pour la classification et l'analyse prédictive. La régression logistique estime la probabilité qu'un événement se produise, tel que voter ou ne pas voter, sur la base d'un ensemble de données donné de variables indépendantes. Comme le résultat est une probabilité, la variable dépendante est bornée entre 0 et 1.

4.1 Iv-1 Régression logistique avec tbl_regression

Ici, nous allons utiliser la fonction tbl_regression du package gtsummary. tbl_regression prend une regression et permet d'afficher les coefficients d'un modèle statistique avec les intervalles de confiance et les p-valeurs. Ici, la variable à expliquer est la pauvreté(p0), les variables explicatives sont le milieu, le genre du chef de ménage, le niveau d'éducation, taille du ménage,

```
# faire un modèle de régréssion la variable dépendante p0(pauvreté)
#les variables indépendantes: milieu, categorie_hsize
mod<-glm(p0~milieu + hgender + categorie_hhsize,
         data = basev, family = binomial
         )
mod%>%tbl_regression()
```

Caractéristique	log(OR)	95% IC	p-valeur
Milieu residence			

Caractéristique	log(OR)	95% IC	p-valeur
Urbain	—	—	
Rural	0,98	0,95 – 1,0	<0,001
Genre du chef de ménage			
Masculin	—	—	
Féminin	-0,48	-0,53 – -0,44	<0,001
Taille du ménage			
Moins de 4 personnes	—	—	
5 à 9 personnes	1,8	1,6 – 1,9	<0,001
10 à 14 personnes	2,4	2,2 – 2,5	<0,001
15 à 19 personnes	2,7	2,5 – 2,8	<0,001
20 personnes et plus	3,4	3,3 – 3,6	<0,001

4.1.1 Iv-1-1 Avec le paramètre include

Le paramètre include permet de choisir les variables à afficher

```
#
mod%>%tbl_regression(include=c(milieu))
```

Caractéristique	log(OR)	95% IC	p-valeur
Milieu residence			
Urbain	—	—	
Rural	0,98	0,95 – 1,0	<0,001

4.1.2 Iv-1-2 Exponentiation des coefficients

Pour une regression logistique il est d'usage d'utiliser d'afficher l'exponentiation des coefficients, ce que l'on peut faire en indiquant **exponentiate=True**

```
mod%>%tbl_regression(exponentiate = TRUE)
```

Caractéristique	OR	95% IC	p-valeur
Milieu residence			
Urbain	—	—	
Rural	2,67	2,58 – 2,76	<0,001
Genre du chef de ménage			
Masculin	—	—	
Féminin	0,62	0,59 – 0,64	<0,001
Taille du ménage			
Moins de 4 personnes	—	—	
5 à 9 personnes	5,77	4,99 – 6,71	<0,001
10 à 14 personnes	10,8	9,36 – 12,6	<0,001
15 à 19 personnes	14,5	12,5 – 16,9	<0,001
20 personnes et plus	31,1	26,8 – 36,2	<0,001

4.1.3 Iv-1-3 Afficher les étoiles de significations

La fonction **add_significance_stars** ajoute des étoiles de significativité à coté des coefficients. Les options hide_ci, hide_p, hide_se permettent de masquer/afficher les intervalles de confiances, les pvaleurs et les écarts types.

```
mod%>%tbl_regression()%>%
  add_significance_stars(hide_ci = FALSE,
                         hide_p = FALSE, hide_se = TRUE)
```

Caractéristique	log(OR)	95% IC	p-valeur
Milieu residence			
Urbain	—	—	
Rural	0,98***	0,95 – 1,0	<0,001
Genre du chef de ménage			
Masculin	—	—	
Féminin	-0,48***	-0,53 – -0,44	<0,001
Taille du ménage			
Moins de 4 persones	—	—	
5 à 9 persones	1,8***	1,6 – 1,9	<0,001
10 à 14 persones	2,4***	2,2 – 2,5	<0,001
15 à 19 persones	2,7***	2,5 – 2,8	<0,001
20 persones et plus	3,4***	3,3 – 3,6	<0,001

4.2 Iv-2 Régression univariée multiple avec tbl_uvregression

La fonction **tbl_uvregression** est utile quand on veut effectuer plusieurs régression univariée. Il faut lui passer un tableau ne contenant que la variable à expliquer et les variables explicatives. La variable à expliquer sera indiqué avec **y**. L'argument method indique la fonction à utiliser pour le calcul des modèles univariés, par exemple glm pour une régression logistique ordinale. On pourra indiquer des paramètres à transmettre à cette fonction avec method.args , par exemple list(family = binomial) dans le cadre d'une régreession logistique binaire.

```
tbl_uni <- tbl_uvregression(
  basev%>%select(p0,milieu, hgender,categorie_hhsize),
  method = glm,
  y=p0,
  method.args = list(family=binomial),
  exponentiate = TRUE,
  hide_n = TRUE)
tbl_uni
```

Caractéristique	OR	95% IC	p-valeur
Milieu residence			
Urbain	—	—	
Rural	2,96	2,87 – 3,06	<0,001
Genre du chef de ménage			
Masculin	—	—	
Féminin	0,40	0,38 – 0,41	<0,001
Taille du ménage			
Moins de 4 persones	—	—	
5 à 9 persones	6,66	5,77 – 7,73	<0,001
10 à 14 persones	12,7	11,0 – 14,7	<0,001
15 à 19 persones	17,8	15,4 – 20,7	<0,001
20 persones et plus	38,4	33,2 – 44,8	<0,001

4.3 Iv-3 Application: regression binomiale

Dans cette partie nous avons effectué une regression logit. La variable indépendante est la pauvreté. Elle est décrit par plusieurs variables: milieu sexe du chef de ménage... **NB**: Nous avons changé les labels des noms des entêtes. On peut trouver le nom des entêtes par show_header_names(). L'intercept est par défaut masqué. On peut l'afficher par intercept=TRUE. On peut ajouter également les valeurs propres globales en cas de besoin par add_global_p et garder les valeurs propres des modalités par keep=TRUE. Bold_labels permet de mettre en gras. **as_gt** permet de transformer en tableau gt. Il faut transformer le tableau en tableau gt pour pouvoir appliquer le titre et la source de donnée. On peut également modifier la note de table avec modify_footnote. Label_number permet de mettre en forme les coefficients de l'odds ratio. label_pvalue met en forme la pvalue

```
mod <- glm(p0~milieu + hgender + categorie_hhsize,
         data = basev, family = binomial)

tbl_mod_b <- mod%>%
  tbl_regression(exponentiate = TRUE,
                 intercept = TRUE,
                estimate_fun = scales::label_number(accuracy=.001,
                                                    decimal.mark ="," ),
                   pvalue_fun= scales::label_pvalue(accuracy=.001,
                                                    decimal.mark=","))%>%
  modify_header(c(label~"**Variables**",estimate~"**Odds
                  ratio**",std.error="**standart error**",
                  p.value ~ "*Test de comparaison* (p-valeur)"))%>%
  modify_footnote(everything()~NA, abbreviation = TRUE)%>%
  add_significance_stars(hide_ci = TRUE, hide_p = FALSE, hide_se = FALSE)%>%
  bold_labels()%>%
  italicize_levels()
```

```
tbl_desc<-basev%>%tbl_custom_summary(
  include = c(milieu, hgender, categorie_hhsize),
  stat_fns = ~proportion_summary(variable="p0",value="pauvre"),
  statistic = ~"{prop}",
  digits = ~ list(
      function(x) {
        style_percent(x, digits = 1)
      },
      0, 0, style_percent, style_percent
      # Mettre les proportion en format %
    )
)%>%
  modify_header(stat_0~"**proportion**")%>%
  modify_footnote(everything()~NA)
```

```
tbl_merge(list(tbl_desc,tbl_mod_b),tab_spanner = c("**Statistique
                                                    descriptive**","**Modèle logit**"))%>%
  as_gt()%>%
  gt::tab_header(
    title=gt::md("**Tableau: Resultat du modèle logistique**"))%>%
  gt::tab_source_note("EHCVM, calculs de l'auteur")
```

Tableau: Resultat du modèle logistique

	Statistique descriptive	Modèle logit

Caractéristique	**proportion**	**Odds ratio**[1]	**standart error**	*Test de comparaison* (p-val
Milieu residence				
Urbain	29,4	—	—	
Rural	55,2	2,671***	0,018	<0,001
Genre du chef de ménage				
Masculin	46,8	—	—	
Féminin	25,8	0,616***	0,023	<0,001
Taille du ménage				
Moins de 4 persones	5,70	—	—	
5 à 9 persones	28,7	5,771***	0,075	<0,001
10 à 14 persones	43,3	10,814***	0,075	<0,001
15 à 19 persones	51,8	14,509***	0,077	<0,001
20 persones et plus	69,9	31,058***	0,077	<0,001
(Intercept)		0,047***	0,075	<0,001

[1] *p<0.05; **p<0.01; ***p<0.001
EHCVM, calculs de l'auteur

5 V Résumé

Les fonctions les plus utilisées sont :

- **tbl_summary**

Argument	Description
`label=`	specify the variable labels printed in table
`type=`	specify the variable type (e.g., continuous, categorical, etc.)
`statistic=`	change the summary statistics presented
`digits=`	number of digits the summary statistics will be rounded to
`missing=`	whether to display a row with the number of missing observations
`missing_text=`	text label for the missing number row
`sort=`	change the sorting of categorical levels by frequency
`percent=`	print column, row, or cell percentages
`include=`	list of variables to include in summary table

- **add...()**

Function	Description
`add_p()`	add *p*-values to the output comparing values across groups
`add_overall()`	add a column with overall summary statistics
`add_n()`	add a column with N (or N missing) for each variable
`add_difference()`	add column for difference between two group, confidence interval, and *p*-value
`add_stat_label()`	add label for the summary statistics shown in each row
`add_stat()`	generic function to add a column with user-defined values
`add_q()`	add a column of *q*-values to control for multiple comparisons

- **format tableau**
- **Exportation** : gtsave , flextable::save_as_docx Les principaux tableaux sortis sont dans le fichier **tableaux_** du dossier **output**. Vous pouvez vous y référer pour les tableaux débordant ou également au fichier html.

Function	Description
`modify_header()`	update column headers
`modify_footnote()`	update column footnote
`modify_spanning_header()`	update spanning headers
`modify_caption()`	update table caption/title
`bold_labels()`	bold variable labels
`bold_levels()`	bold variable levels
`italicize_labels()`	italicize variable labels
`italicize_levels()`	italicize variable levels
`bold_p()`	bold significant *p*-values

6 VI Bibliographie et webographie

- *https://www.danieldsjoberg.com/gtsummary-weill-cornell-presentation/#59*
- *https://github.com/ddsjoberg/gtsummary*
- *https://www.danieldsjoberg.com/gtsummary/*
- *Reproducible Summary Tables with the gtsummary Package by Daniel D. Sjoberg, Karissa Whiting, Michael Curry, Jessica A. Lavery, Joseph Larmarange*
- *https://larmarange.github.io/analyse-R/gtsummary.html*