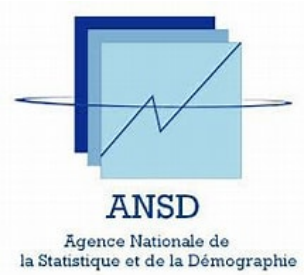
RÉPUBLIQUE DU SÉNÉGAL

Un Peuple - Un But - Une Foi



Agence Nationale de la Statistique et de la Démographie

École Nationale de la Statistique et de l'Analyse Économique

A LA DECOUVERTE DE GTSUMMARY

Illustration avec la base EHCVM 2021 de la Côte d'Ivoire

Rédigé par :

NGUEMFOUO NGOUMTSA Célina

Élève Ingénieure Statisticienne Économiste

Sous la supervision de :

M. Aboubacar HEMA

Data analyst

Année scolaire : 2024/2025

Avant-propos

Dans un monde en constante évolution, le rôle des Ingénieurs Statisticiens Économistes (ISE) est devenu indispensable pour répondre aux enjeux complexes des sociétés modernes. Leur expertise en matière de collecte, d'analyse et d'interprétation des données permet aux décideurs publics et privés de mieux comprendre les dynamiques économiques, sociales et démographiques, et de prendre des décisions éclairées. Grâce à leur formation polyvalente, les ISE contribuent de manière décisive à l'élaboration et à l'évaluation des politiques publiques, ainsi qu'au développement de stratégies économiques efficaces.

C'est dans ce cadre que l'École nationale de la Statistique et de l'Analyse économique Pierre Ndiaye (ENSAE) du Sénégal s'est imposée comme l'un des principaux établissements de formation en Afrique francophone. L'ENSAE offre plusieurs parcours de formation spécialisés dans la statistique et l'économie appliquée, avec des filières comme :

- **Les Analystes Statisticiens (AS)** : une formation de trois ans, qui forme des techniciens en statistique capables de traiter et d'analyser des données à des fins variées.
- **Les Ingénieurs Statisticiens Économistes (ISE)** : un cycle court de trois ans, et un cycle long de cinq ans, qui offre une formation approfondie à ceux qui souhaitent devenir des cadres spécialisés dans l'analyse statistique et économique.

L'accès à l'ENSAE est soumis à un concours rigoureux ouvert aux étudiants détenteurs du baccalauréat ou d'un diplôme universitaire selon le niveau d'admission. De plus, l'école propose des bourses pour les étudiants méritants, facilitant ainsi l'accès à une formation d'excellence, tant pour les étudiants sénégalais que pour ceux venant d'autres pays africains francophones. En tant que membre du Réseau des Écoles Africaines de la Statistique, l'ENSAE collabore étroitement avec d'autres institutions prestigieuses comme l'Institut Sous-Régional de Statistique et d'Économie Appliquée (ISSEA) au Cameroun et l'École Nationale Supérieure de Statistique et d'Économie Appliquée (ENSEA) en Côte d'Ivoire. Ce réseau permet d'harmoniser les programmes de formation et d'offrir aux élèves une reconnaissance internationale de leurs compétences.

Durant leur cursus, les étudiants de la filière ISE réalisent de nombreux rapports pour leur permettre d'appliquer les concepts et de se familiariser avec leurs outils de travail. Le présent rapport s'inscrit dans cette démarche, et a pour thème **« A la découverte du package gtsummary : Illustration avec la base EHCVM 2021 de la Côte d'Ivoire »**.

Sommaire

Contents

Résumé

Ce rapport explore l'utilisation du package gtsummary dans le cadre de l'analyse statistique descriptive, en s'appuyant sur les données de l'EHCVM 2021 de la Côte d'Ivoire. L'objectif est de montrer comment cet outil facilite la génération de tableaux synthétiques et permet une visualisation claire des statistiques descriptives. Après une présentation du package et des données, nous illustrons les fonctionnalités de la fonction tbl_summary, qui permet de résumer efficacement les variables d'un jeu de données. À travers des exemples concrets, nous mettons en évidence la simplicité et la puissance de gtsummary, notamment pour automatiser la production de rapports statistiques. Ce travail s'inscrit dans une démarche plus large d'amélioration des outils d'analyse de données, essentielle pour les statisticiens et économistes en quête d'efficacité et de reproductibilité.

Introduction

Dans le cadre de l'analyse des données, les ingénieurs statisticiens économistes doivent extraire rapidement des informations pertinentes à partir de vastes ensembles de données. L'une des premières étapes exploratoires consiste à générer des résumés statistiques (summary), permettant d'identifier les tendances générales, et d'obtenir une vision synthétique des données avant d'engager des analyses plus approfondies. Cette étape facilite la compréhension des distributions, la détection des erreurs potentielles et l'adaptation des méthodes d'analyse. Les fonctions classiques de R, comme *summary()*, permettent d'obtenir des statistiques descriptives de base. Cependant, elles présentent plusieurs limites :

- Elles sont souvent générales et ne permettent pas d'obtenir des résumés adaptés à des types de données spécifiques (textes, facteurs, valeurs manquantes).
- Elles manquent de flexibilité pour afficher des statistiques avancées comme l'asymétrie, l'aplatissement, ou des quantiles personnalisés.
- Leur présentation n'est pas toujours optimisée pour l'interprétation et l'exportation des résultats.

Ce rapport vise à explorer le package **gtsummary**, qui fournit des fonctions plus complètes et adaptées pour générer des résumés statistiques avancés. Nous allons détailler son fonctionnement et illustrer ses principales fonctionnalités à travers des exemples concrets. Afin de repondre à cet objectif, notre rapport s'articulera en 2 chapitres. Le premier fera une brève présentation du package **gtsummary** et de l'**EHCVM**, et le second s'attardera en particulier sur la fonction **tbl_summary** du package.

Chapitre 1 : présentation du package et de l'EHCVM

I. Présentation du package gtsummary

Le package **gtsummary** est une extension de R conçue pour faciliter la création de tableaux de synthèse et de rapports statistiques bien structurés. Il est particulièrement utilisé en biostatistique et en analyse de données pour produire des résumés descriptifs, des comparaisons de groupes et des résultats de modèles statistiques sous un format clair et lisible. Développé pour être compatible avec des outils de présentation comme gt, flextable et kableExtra, gtsummary permet d'obtenir des tableaux prêts à être exportés dans des documents professionnels (Word, PDF, HTML). Il s'intègre parfaitement avec les pipelines de tidyverse, notamment dplyr, facilitant ainsi l'automatisation des analyses.

L'objectif principal de gtsummary est de simplifier la génération de tableaux statistiques tout en garantissant une présentation soignée et adaptée aux rapports scientifiques, une compatibilité avec plusieurs formats d'exportation (HTML, Word, LaTeX), ainsi qu'une intégration fluide avec dplyr pour un usage efficace dans les flux de travail en R.

Le package gtsummary propose plusieurs fonctions clés :

- **tbl_summary() :** Génère un tableau de statistiques descriptives pour un jeu de données.
- **tbl_regression() :** Produit des tableaux de résultats pour les modèles de régression (linéaire, logistique, etc.).
- **tbl_compare() :** Compare des statistiques entre plusieurs groupes.
- **tbl_merge() :** Fusionne plusieurs tableaux gtsummary.
- **as_flextable() et as_gt() :** Convertit les tableaux pour une personnalisation avancée et une meilleure exportation.

Dans la suite du rapport, nous nous concentrerons sur la fonction **tbl_summary** et de ses paramètres. Pour l'illusration, nous utiliserons les bases *ménages* et *welfare* de l'EHCVM en Côte d'Ivoire.

II. Présentation de l'EHCVM

L'**Enquête Harmonisée sur les Conditions de Vie des Ménages** (EHCVM) est une initiative régionale menée dans plusieurs pays de l'UEMOA, dont la **Côte d'Ivoire**, dans le but de produire des données comparables sur le bien-être des ménages. Cette enquête s'inscrit dans un cadre d'harmonisation des statistiques de la pauvreté et des conditions de vie afin d'améliorer l'élaboration et l'évaluation des politiques publiques.

L'EHCVM vise à fournir des informations détaillées sur :

- **Les conditions de vie des ménages** : accès aux services de base, logement, éducation, emploi.
- **La consommation et la pauvreté** : estimation des dépenses des ménages pour calculer les indicateurs de bien-être économique.
- **Les inégalités et la vulnérabilité** : analyse des écarts entre différents groupes sociaux et géographiques.
- **L'impact des politiques publiques** : suivi des stratégies nationales de lutte contre la pauvreté et le développement humain.

L'enquête repose sur un échantillonnage représentatif au niveau national, couvrant les zones urbaines et rurales. Les données sont collectées à l'aide de questionnaires détaillés adressés aux ménages, permettant d'analyser le profil sociodémographique des ménages, leur accès aux infrastructures et services sociaux, leurs revenus, dépenses et stratégies de subsistance.

Parmi les différentes bases de l'EHCVM, deux bases essentielles seront utilisées dans cette analyse :

- **La base "ménage"** : Contient les informations générales sur les ménages (composition, logement, accès aux services).
- **La base "welfare"** : Fournit des indicateurs clés sur la consommation, la pauvreté et le bien-être économique des ménages. Ces bases de données permettent une exploration approfondie des dynamiques socioéconomiques en Côte d'Ivoire et constituent des sources précieuses pour les analyses quantitatives.

Chapitre 2 : fontion tbl summary

I. Base ménage

la base ménage a d'abord été importée. Pour ce faire, le package **haven** a été utilisé. Il s'agit d'un fichier .dta.

```
library(haven)
menage <- read_dta(
  "CIV_2021_EHCVM_Stata/ehcvm_menage_civ2021.dta")
```

Voici les premières lignes de la base

```
library(knitr)
kable(head(menage[, 1:10]))
```

country	hhid	grappe	menage	vague	logem	mur	toit	sol	eauboi_ss
	101	NA	NA	NA	NA	NA	NA	NA	NA
CIV	102	1	2	1	3	1	1	1	1
CIV	103	1	3	1	3	1	1	1	1
CIV	104	1	4	1	4	1	1	1	1
CIV	105	1	5	1	3	1	1	1	1
CIV	106	1	6	1	3	1	1	1	1

Visualisons à présent un summary des variables *logem, toit* et *sol*.

Pour cela, il faut charger la librairie **gtsummary**.

```
library(gtsummary)
suppressMessages(
  menage %>% select(logem,toit,sol) %>% tbl_summary()
  )
```

Characteristic	N = 13,863[1]
Occupation logement	
1	2,844 (22%)
2	4,508 (35%)
3	2,702 (21%)
4	2,906 (22%)
9	5 (<0.1%)
Unknown	898
toit en materiaux definitifs	
0	1,337 (10%)
1	11,628 (90%)
Unknown	898
Sol en materiaux definitifs	
0	1,736 (13%)
1	11,229 (87%)
Unknown	898

[1] n (%)

Affichons à présent les modalités des variables:

```
menage %>%
  labelled::to_factor() %>%
  select(logem,toit,sol) %>%
  tbl_summary()
```

Characteristic	**N = 13,863**[1]
Occupation logement	
Proprietaire titre	2,844 (22%)
Proprietaire sans titre	4,508 (35%)
Locataire	2,702 (21%)
Autre	2,906 (22%)
9	5 (<0.1%)
Unknown	898
toit en materiaux definitifs	
Non	1,337 (10%)
Oui	11,628 (90%)
Unknown	898
Sol en materiaux definitifs	
Non	1,736 (13%)
Oui	11,229 (87%)
Unknown	898

[1] n (%)

Attribuons à présent des labels aux différentes variables sélectionnées

```
menage %>%
  labelled::to_factor() %>%
  select(logem,toit,sol) %>%
  tbl_summary(
    label=list(logem~"Type de logement du CM",
               toit~"Type de toit de la maison du CM",
               sol~"Type de sol de la maison du CM"))
```

Characteristic	**N = 13,863**[1]
Type de logement du CM	
Proprietaire titre	2,844 (22%)
Proprietaire sans titre	4,508 (35%)
Locataire	2,702 (21%)
Autre	2,906 (22%)
9	5 (<0.1%)
Unknown	898
Type de toit de la maison du CM	
Non	1,337 (10%)
Oui	11,628 (90%)
Unknown	898
Ype de sol de la maison du CM	
Non	1,736 (13%)
Oui	11,229 (87%)
Unknown	898

[1] n (%)

Modifions à présent le titre du tableau

```
menage %>% labelled::to_factor() %>%
  select(logem,toit,sol) %>%
  tbl_summary(
    label=list(logem~"Type de logement du CM",
               toit~"Type de toit de la maison du CM",
               sol~"Type de sol de la maison du CM")) %>%
  modify_header(label="Caractéristiques de l'habitat du CM")
```

Considérons à présent les variables *superf, grosrum* et *petitrum* Comme il s'agit de variable numériques, le paramètre **statistic** a été ajouté pour visualiser la **moyenne** (mean) et l'écartype (sd).

Caractéristiques de l'habitat du CM	**N = 13,863**[1]
Type de logement du CM	
Proprietaire titre	2,844 (22%)
Proprietaire sans titre	4,508 (35%)
Locataire	2,702 (21%)
Autre	2,906 (22%)
9	5 (<0.1%)
Unknown	898
Type de toit de la maison du CM	
Non	1,337 (10%)
Oui	11,628 (90%)
Unknown	898
Type de sol de la maison du CM	
Non	1,736 (13%)
Oui	11,229 (87%)
Unknown	898

[1] n (%)

```
menage %>%
  labelled::to_factor() %>%
  select(superf,grosrum,petitrum) %>%
  tbl_summary(
    label=list(superf~"Superficie agricole en moyenne par ménage",
               grosrum~"Nombre de gros ruminants en moyenne par ménac
               petitrum~"Nombre de petits ruminants en moyenne par me
    statistic = list(superf~"{mean}({sd})",
                     grosrum~"{mean}({sd})",
                     petitrum~"{mean}({sd})")) %>%
  modify_header(label="Caractéristiques de l'habitat du CM")
```

les valeurs présentes dans le tableau comme indiqué dans la dernière ligne sont les moyennes

Caractéristiques de l'habitat du CM	**N = 13,863**[1]
Superficie agricole en moyenne par ménage	24,660,570(1,441,480,913)
Unknown	898
Nombre de gros ruminants en moyenne par ménage	0.7405(9.1305)
Unknown	898
Nombre de petits ruminants en moyenne par ménage	0.93(3.83)
Unknown	898

[1]Mean(SD)

et entre parenthèse l'écart-type pour chacune des variables. On constate la présence de valeurs manquantes.

Le code suivant permet de toujours les afficher:

```
menage %>%
  labelled::to_factor() %>%
  select(superf,grosrum,petitrum) %>%
  tbl_summary(
    label=list(superf~"Superficie agricole en moyenne par ménage",
               grosrum~"Nombre de gros ruminants en moyenne par ménac
               petitrum~"Nombre de petits ruminants en moyenne par me
    statistic = list(superf~"{mean}({sd})",
                     grosrum~"{mean}({sd})",
                     petitrum~"{mean}({sd})"),
    missing="always") %>%
  modify_header(label="Caractéristiques de l'habitat du CM")
```

On a également la possibilté de renommer les valeurs manquantes.

```
menage %>%
  labelled::to_factor() %>%
  select(superf,grosrum,petitrum) %>%
  tbl_summary(
    label=list(superf~"Superficie agricole en moyenne par ménage",
               grosrum~"Nombre de gros ruminants en moyenne par ménac
```

Caractéristiques de l'habitat du CM	**N = 13,863**[1]
Superficie agricole en moyenne par ménage	24,660,570(1,441,480,913)
Unknown	898
Nombre de gros ruminants en moyenne par ménage	0.7405(9.1305)
Unknown	898
Nombre de petits ruminants en moyenne par ménage	0.93(3.83)
Unknown	898

[1]Mean(SD)

Caractéristiques de l'habitat du CM	**N = 13,863**[1]
Superficie agricole en moyenne par ménage	24,660,570(1,441,480,913)
Valeurs manquantes	898
Nombre de gros ruminants en moyenne par ménage	0.7405(9.1305)
Valeurs manquantes	898
Nombre de petits ruminants en moyenne par ménage	0.93(3.83)
Valeurs manquantes	898

[1]Mean(SD)

```
                petitrum~"Nombre de petits ruminants en moyenne par me
    statistic = list(superf~"{mean}({sd})",
                     grosrum~"{mean}({sd})",
                     petitrum~"{mean}({sd})"),
    missing="always",
    missing_text = "Valeurs manquantes") %>%
  modify_header(label="Caractéristiques de l'habitat du CM")
```

On peut en outre choisir le nombre de chiffres après la virgule. Choisissons aucun chiffre après la virgule :

Caractéristiques de l'habitat du CM	**N = 13,863**[1]
Superficie agricole en moyenne par ménage	24,660,570(1,441,480,913)
Valeurs manquantes	898
Nombre de gros ruminants en moyenne par ménage	1(9)
Valeurs manquantes	898
Nombre de petits ruminants en moyenne par ménage	1(4)
Valeurs manquantes	898

[1]Mean(SD)

```
menage %>%
  labelled::to_factor() %>%
  select(superf,grosrum,petitrum) %>%
  tbl_summary(
    label=list(superf~"Superficie agricole en moyenne par ménage",
               grosrum~"Nombre de gros ruminants en moyenne par ménag
               petitrum~"Nombre de petits ruminants en moyenne par m
    statistic = list(superf~"{mean}({sd})",
                     grosrum~"{mean}({sd})",
                     petitrum~"{mean}({sd})"),
    digits = everything()~c(0,0,0),
    missing="always",
    missing_text = "Valeurs manquantes") %>%
  modify_header(label="Caractéristiques de l'habitat du CM")
```

II. Base welfare

Commençons par charger la base en question.

```
library(haven)
welfare <- read_dta(
  "CIV_2021_EHCVM_Stata/ehcvm_welfare_civ2021.dta"
  )
```

voici un aperçu de la base chargée :

```
library(knitr)
kable(head(welfare[, 1:10]))
```

grappe	menage	country	year	hhid	vague	month	zae	region	milieu
1	11	CIV	2021	111	1	2022-01-01	6	1	1
1	27	CIV	2021	127	1	2022-02-01	6	1	1
1	7	CIV	2021	107	1	2022-01-01	6	1	1
1	8	CIV	2021	108	1	2022-01-01	6	1	1
1	10	CIV	2021	110	1	2022-01-01	6	1	1
1	9	CIV	2021	109	1	2022-01-01	6	1	1

Faisons à présent un summary pour les variables hgender, hage, hmstat, heduc, hdiploma. Seule la variable age est numérique, donc c'est uniquement pour elle que le paramètre statistic (mean et sd) a été défini. Les modalités ont été affichées, le titre du tableau renomé, le nombre de chiffre après la virgule fixé, et les valeurs manquantes affichées. Voici le résultat:

```
welfare %>%
  labelled::to_factor() %>%
```

```
  select(hgender, hage, hmstat, heduc, hdiploma) %>%
  tbl_summary(label=list(hgender~"Genre du CM",
                         hage~"Âge du CM",
                         hmstat~"Situation matrimoniale du CM",
                         heduc~"Niveau d'étude du CM",
                         hdiploma~"Diplome le plus élevé"),
              statistic = list(hage~"{mean}({sd})"),
              digits = everything()~c(0,0,0),
              missing="always",
              missing_text = "Valeurs manquantes")%>%
  modify_header(label="Caractéristiques du chef de ménage")
```

Conclusion

L'utilisation du package gtsummary s'avère être une solution efficace pour la synthèse et la présentation des statistiques descriptives, particulièrement dans le cadre de l'analyse d'enquêtes comme l'EHCVM 2021. Grâce à ses fonctionnalités intuitives, il permet de gagner du temps et d'assurer une meilleure lisibilité des résultats. L'application de tbl_summary démontre qu'il est possible de produire des tableaux complets et personnalisables avec un minimum d'effort, tout en garantissant une présentation standardisée. Ce rapport met ainsi en avant l'intérêt d'intégrer des outils comme gtsummary dans les pratiques courantes des statisticiens, contribuant ainsi à une analyse plus rapide et plus fiable des données.

Table de matières

Caractéristiques du chef de ménage	**N = 12,965**[1]
Genre du CM	
Masculin	10,689 (82%)
Féminin	2,276 (18%)
Valeurs manquantes	0
Âge du CM	46(14)
Valeurs manquantes	0
Situation matrimoniale du CM	
Célibataire	1,907 (15%)
Marié(e) monogame	7,171 (55%)
Marié(e) polygame	1,656 (13%)
Union libre	811 (6%)
Veuf(ve)	1,078 (8%)
Divorcé(e)	161 (1%)
Séparé	181 (1%)
Valeurs manquantes	0
Niveau d'étude du CM	
Aucun	7,444 (57%)
Maternelle	3 (0%)
Primaire	2,544 (20%)
Second. gl 1	1,409 (11%)
Second. tech. 1	36 (0%)
Second. gl 2	791 (6%)
Second. tech. 2	73 (1%)
Postsecondaire	257 (2%)
Superieur	407 (3%)
Valeurs manquantes	1
Diplome le plus élevé	
Aucun	9,759 (75%)
cepe	1,499 (12%)
bepc	755 (6%)
cap	35 (0%)
bt	41 (0%)