**Test de management et d’analyse des données avec R**

**Enseignant: Aboubacar Hema**

**Instructions :**

Vous avez 3 heures à partir du moment où vous recevez ce test pour envoyer vos réponses dans le compte GitHub du cours ([Abson-dev/Projet-statistique-sous-R](https://github.com/Abson-dev/Projet-statistique-sous-R)). Les soumissions tardives ne seront pas prises en compte. Veuillez renvoyer dans un fichier zip :

1. le document .log
2. le do-file commenté .do

Vous devez compléter ce test **seul, sans l’aide d’une tierce personne**. En revanche, l’utilisation d’internet est **autorisée pour effectuer des recherches documentaires ou consulter le support R**.

Nous évaluerons vos compétences dans cinq catégories :

* Compétences de base en management des données avec R
* Compétences avancées en management des données avec R
* Compétences en statistiques
* Clarté des codes et reproductibilité des résultats
* Interprétation des résultats
* Compétences rédactionnelles

Pour tous les exercices, **commentez** vos codes et les sorties.

**Epreuve :**

Cette épreuve est composée de 2 parties ;

1. Gestion et nettoyage des bases de données,
2. Empilement et Fusion des données

Les bases de données utilisées dans ce test proviennent d’enquêtes réalisées auprès des mères pour analyser la composition nutritionnelle du régime alimentaire des mères et de leurs enfants au cours des dernières 24 heures. Les données ont été collectées à deux moments distincts : au début de l’étude (Baseline) et à la fin de l’étude (Endline). Elles comprennent :

1. Les informations sur l’alimentation des mères lors des enquêtes Baseline (***food\_comp\_mother\_baseline.dta***) et Endline (***food\_comp\_mother\_endline.dta***). Ces bases détaillent les quantités d’énergie et de nutriments (protéines, lipides, calcium, fer, zinc, vitamine B6, vitamine B12, vitamine C) consommées lors de chacun des 4 repas (petit-déjeuner, déjeuner, diner, snacks/gouter).
2. Les données sur l’alimentation des enfants recueillies aux mêmes périodes : Baseline (***food\_comp\_child\_baseline.dta***) et Endline (***food\_comp\_child\_endline.dta***), données organisées selon le même modèle que pour les mères.
3. Les caractéristiques des ménages ayant participé à l’étude : ***base\_menage.dta***

Votre tâche est de réaliser le nettoyage, la fusion et l’analyse des données en suivant les étapes décrites ci-dessous. Écrivez avec commentaires les **commandes Stata** correspondantes pour chaque étape.

**Partie 1 : Gestion et nettoyage des bases de données**

1. Vérifiez la présence de doublons dans les bases de données Baseline, Endline et ménage. Supprimez les doublons si nécessaire.
2. Assurez-vous que les noms des variables sont cohérents entre les bases de données Baseline et Endline (par exemple, les noms des nutriments doivent être identiques). Si ce n’est pas le cas, apportez des corrections.
3. Certaines variables peuvent contenir des données manquantes, cependant suivant une logique bien établie, il est possible de corriger ces données manquantes. Veuillez vérifier soigneusement les données et corriger les données manquantes de certaines variables si possibles.
4. Quelle est la consommation d’énergie moyenne à chaque repas pour l’ensemble des mères lors de l’enquête de base ?
5. Après vérification et correction de chaque base de données :
   * 1. Sauvegarder les changements de la base ***food\_comp\_mother\_baseline.dta***en***mother\_baseline\_v1.dta***
     2. Sauvegarder les changements de la base ***food\_comp\_mother\_endline.dta***en***mother\_endline\_v1.dta***
     3. Sauvegarder les changements de la base ***food\_comp\_child\_baseline.dta***en***child\_baseline\_v1.dta***
     4. Sauvegarder les changements de la base ***food\_comp\_child\_endline.dta***en***child\_endline\_v1.dta***
     5. Sauvegarder les changements de la base ***base\_menage.dta*** en ***base\_menage\_final.dta***

**Partie 2 : Empilement et Fusion des données**

Maintenant que vous avez vérifié et corrigé les possibles incohérences dans les bases de données suivre les instructions suivantes :

1. **Baseline**
   * 1. Empilez les bases de données des mères (***mother\_baseline\_v1.dta***) et des enfants (***child\_baseline\_v1.dta***) de sorte que pour chaque ménage de l’étude, on retrouve une ou plusieurs lignes concernant la mère et sa consommation journalière et une ou plusieurs lignes concernant la consommation de l’enfant.
     2. Renommez toutes les variables de consommation *energ\_kcal* jusqu’à *vit\_c\_mcg* en ajoutant le suffixe***\_b*** *pour* faire référence à l’enquête Baseline.
     3. Créez une base de données qui résume les consommations journalières totales par individu (somme des 4 repas) pour l’énergie et tous les nutriments en utilisant la commande collapse.
     4. Sauvegardez la base de données finale contenant seulement les *hhid s1\_q2 energ\_kcal protein\_g lipid\_tot\_g calcium\_mg iron\_mg zinc\_mg* sous le nom « ***baseline\_final.dta »****.* Veuillez noter que pour cette base de données, pour chaque ménage, nous avons une ligne pour la mère et une ligne pour l’enfant.
2. **Endline**
   * 1. Empilez les bases de données des mères (***mother\_endline\_v1.dta***) et des enfants (***child\_endline\_v1.dta***) de sorte que pour chaque ménage de l’étude, on retrouve une ou plusieurs lignes concernant la mère et sa consommation journalière et une ou plusieurs lignes concernant la consommation de l’enfant. Renommez toutes les variables de consommation *energ\_kcal* jusqu’à *vit\_c\_mcg* en ajoutant le suffixe***\_e*** *pour* faire référence à l’enquête Endline.
     2. Créez une base de données qui résume les consommations journalières totales par individu (somme des 4 repas) pour l’énergie et tous les nutriments en utilisant la commande collapse.
     3. Sauvegarder la base de données finale contenant seulement les *hhid s1\_q2 energ\_kcal protein\_g lipid\_tot\_g calcium\_mg iron\_mg zinc\_mg* sous le nom « ***endline\_final.dta »****.* Veuillez noter que pour cette base de données, pour chaque ménage, nous avons une ligne pour la mère et une ligne pour l’enfant.
3. Fusionnez les données ***baseline\_final.dta*** et ***endline\_final.dta.*** La base de données résultant de cette fusion doit comporter le même nombre d’observations que les bases ***baseline\_final.dta*** et ***endline\_final.dta.*** Ainsi, sur une ligne d’observation, on retrouve à la fois les informations de l’enquête Baseline et de l’enquête Endline, concernant l’individu (mère ou enfant).
4. En utilisant la base ***base\_menage\_final***, fusionnez les informations sur les caractéristiques des ménages telles que région, commune, village, statut matrimonial, taille du ménage, niveau d’éducation du chef de ménage, le ratio de dépendance, score HFIAS, la variable de traitement **T1**.
5. Sauvegarder la base de données finale sous le nom *base\_finale.*