RWANDA

IFPRI

Introduction to Household Data Analysis

Using Stata to Describe, Transform, and Analyze Data

Kigali, Rwanda

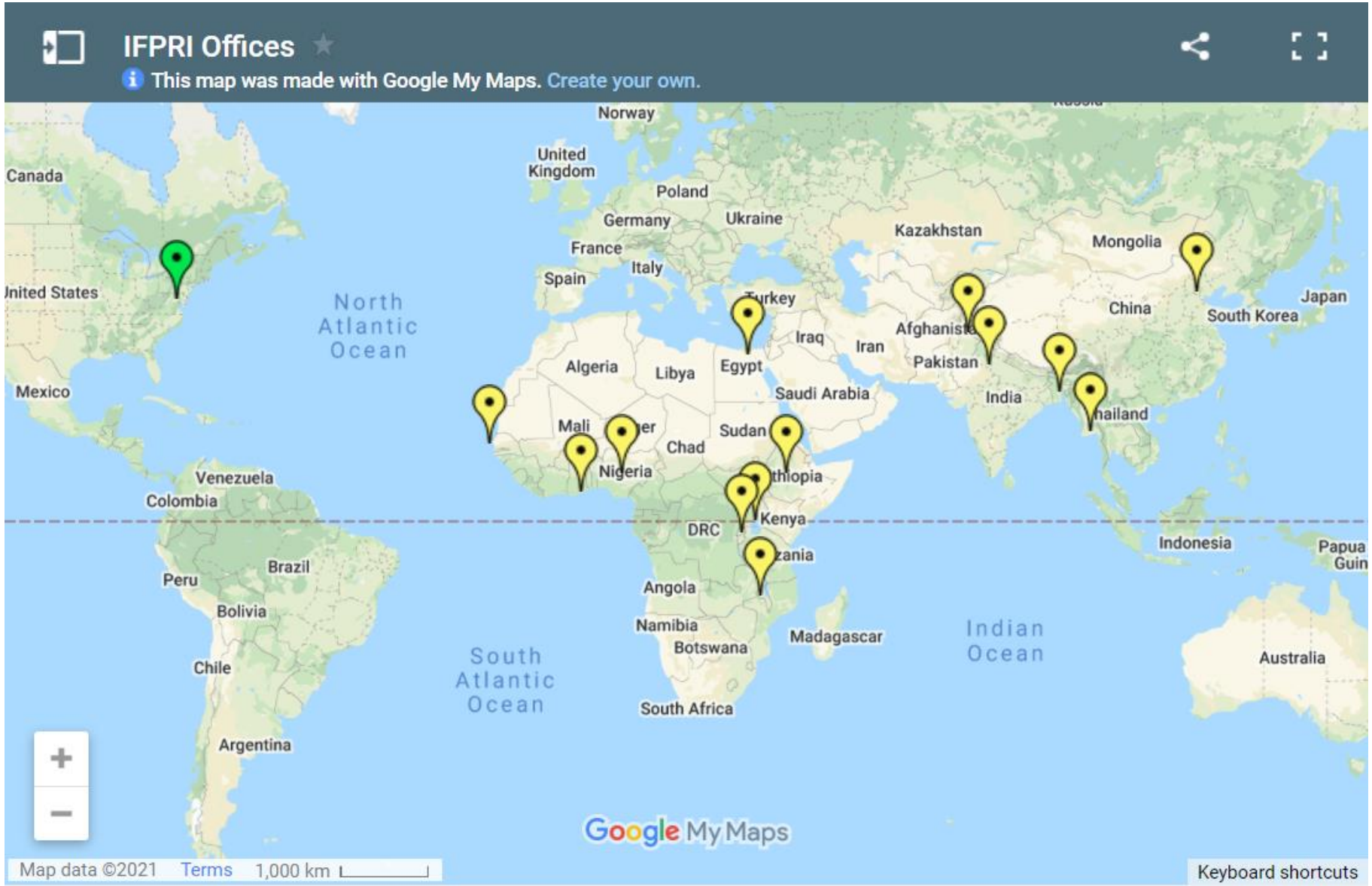
March 1-3, 2022

International Food Policy Research Institute (IFPRI)

Contact: Gracie Rosenbach (g.rosenbach@cgiar.org)

International Food Policy Research Institute

International Food Policy Research Institute



IFPRI Rwanda

- The Rwanda Strategy Support Program (Rwanda SSP) is an initiative aligned with the Government of Rwanda's vision for accelerating agricultural transformation and rural development throughout the country.
  - Collaborative policy analysis
  - Capacity development and knowledge sharing

IFPRI Rwanda

1. How can MINAGRI best allocate their budget to reach their targets?

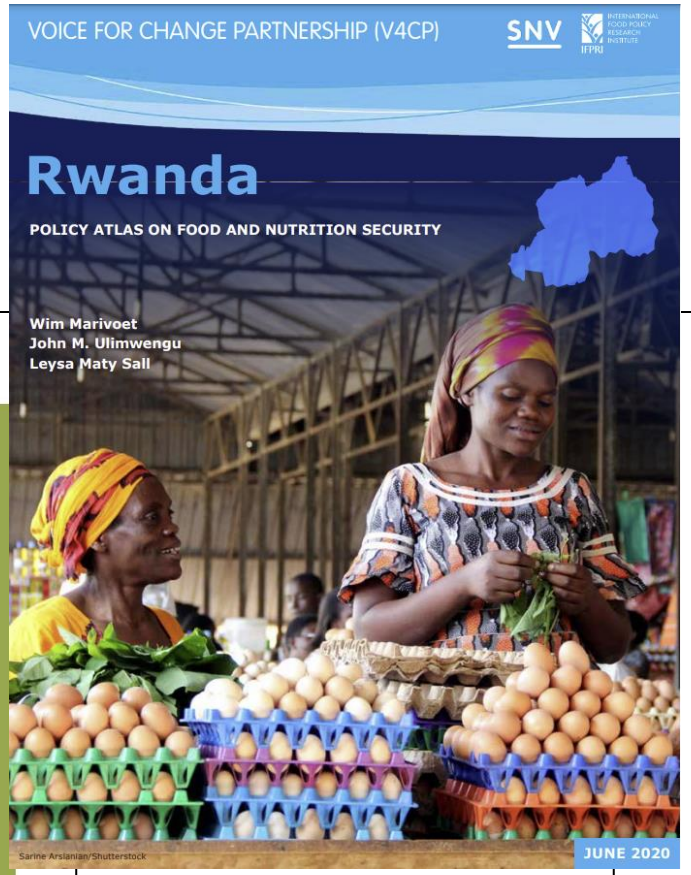
2. How can we sustainably intensify and modernize agriculture?

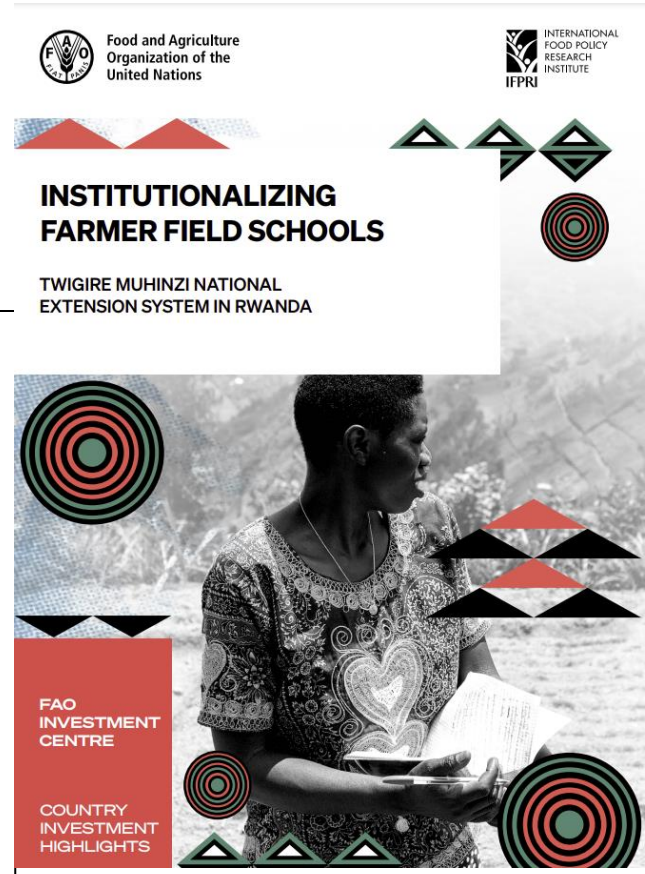
3. How can the agriculture sector best contribute to reaching nutrition goals?

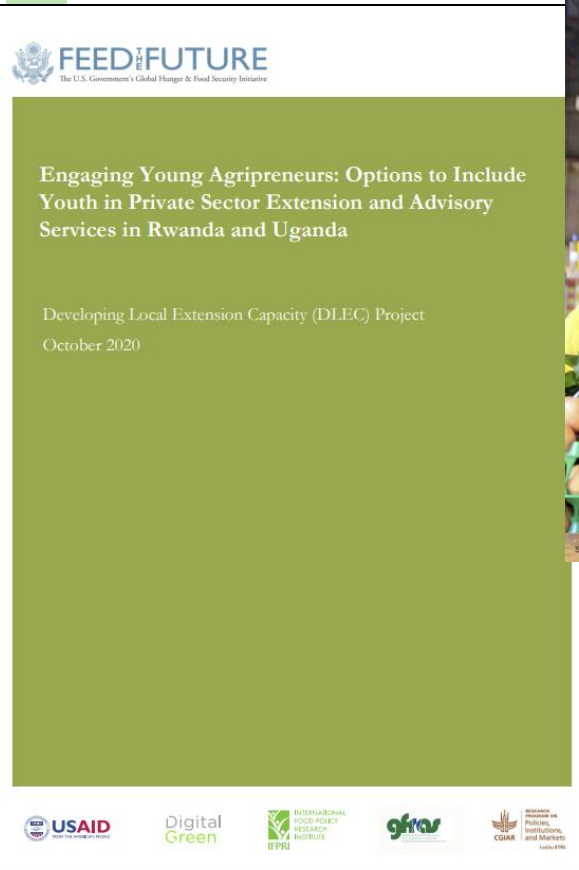
4. How can we strengthen the entire agri-food system (value chains, exports, etc.)?

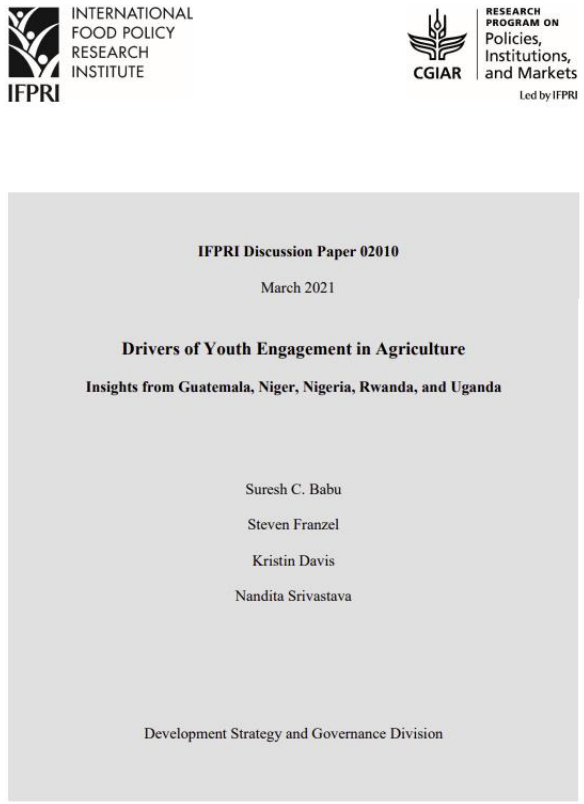
5. How can we ensure agriculture policies are inclusive of everyone (e.g. women, youth, all agri-food system actors along the value chain)?

IFPRI Rwanda









What Is Stata?

- A general-purpose statistical software package
- Mostly used in economics, sociology, political science, biomedicine, and epidemiology

What Is Stata?

- A general-purpose statistical software package
- Mostly used in economics, sociology, political science, biomedicine, and epidemiology
- Stata's capabilities include data management, statistical analysis, graphics, simulations, regression, and custom programming.

Course Overview

- Introduction to the course, introduction to Stata, and describing data
  - *Count, codebook, tabulate, summarize, creating figures, etc.*

Course Overview

- Introduction to the course, introduction to Stata, and describing data
  - *Count, codebook, tabulate, summarize, creating figures, etc.*

- Transforming Data
  - *Creating new variables, labeling variables, collapsing data, merging two datasets, etc.*

Course Overview

- Introduction to the course, introduction to Stata, and describing data
  - *Count, codebook, tabulate, summarize, creating figures, etc.*
- Transforming Data
  - *Creating new variables, labeling variables, collapsing data, merging two datasets, etc.*
- Analyzing Data
  - *T-tests, correlations, regressions, etc.*

Course Objectives

- By the end of this course, you will be able to:
  - Understand how household surveys can help to inform policy decisions
  - Utilize Stata to get a better understanding of data and what it can tell us
  - Communicate findings from the EICV5

Course Objectives

- By the end of this course, you will be able to:
  - Understand how household surveys can help to inform policy decisions
  - Utilize Stata to get a better understanding of data and what it can tell us
  - Communicate findings from the EICV5

- We'll learn how to answer questions such as:
  - What percent of poor households live in rural areas?
  - What percent of households have a female household head?
  - Is household size correlated with household consumption?

Course Data

- Data: Fifth Integrated Household Living Conditions Survey (EICV5), NISR 2016/17

Course Data – EICV 5

- Nationally representative – includes households from each district
- Cross-sectional survey – the main survey does *not* visit the same households in each iteration

Course Data – EICV 5

- Nationally representative – includes households from each district
- Cross-sectional survey – the main survey does *not* visit the same households in each iteration
- Contains modules on:
  - Housing conditions
  - Education
  - Health
  - Income and employment
  - Household consumption

Course Data – EICV 5

- Today, we're going to be looking at the household characteristics data file: cs_S0_S5_Household
- The S0 and S5 corresponds to the sections of the questionnaire that this data represent

Course Data – EICV 5

- Today, we're going to be looking at the household characteristics data file: cs_S0_S5_Household
- The S0 and S5 corresponds to the sections of the questionnaire that this data represent

SECTION 5: HOUSING

REQUIRED: THE HEAD OF THE HOUSEHOLD or the most knowledgeable person
At this point, I would like to ask you some questions concerning your housing. Whereby housing refers to every room and separate structure used by members of your household

PART A: BACKGROUND AND STATUS OF THE HOUSING OCCUPANCY

1	2	3	4	5	6
Type of habitat *Umudugudu (new recommended rural resettlement)*.....1 Unplanned clustered rural housing.....2 Isolated rural housing.....3 Urban informal /unplanned housing area.....4 Old *resettlement*.....5 Modern planned urban area.....6 Other (Specify) 7	Type of Dwelling A single house occupied by one household dwelling.....1 =>Q5 A house occupied by multiple Household s.....2 Multi-storied building with one household.....3 =>Q5 Multi-storied building with more households.....4 =>Q5 Group of enclosed dwellings: multiple households.....5 =>Q5 Group of enclosed dwellings occupied by a single household.....6 =>Q5 Other (Specify) .....7 =>Q5	How many other households share the dwelling with you? Number	Do you share any of the rooms in the dwelling with other households? Yes.....1 No.....2	5. How many rooms does your household occupy? (Excluding bathroom, toilet, kitchen, corridor and stables) **Excluding rooms extensively occupied by other HHs**	6. How many rooms does your household have for sleeping in? **Excluding rooms extensively occupied by other HHs**

Day 1 Data

- Let’s double-click to open file *cs_S0_S5_Household.dta* (found on your desktops)

Practice 1 – count, codebook, and tabulate

- Please practice the following codes in Stata on your own:
  - count
    - *How many observations are in the dataset?*
  - codebook poverty
    - *What does the variable **poverty** tell us?*
  - tab poverty
    - *How many households are non-poor?*

Practice 2 – tabulate, histogram, and summarize

- Please practice the following codes in Stata on your own:
  - tab s5aq5
    - *What percent of households have 3 rooms in their homes?*
  - hist s5aq5
    - *What is the most common number of rooms?*
  - sum s5aq5
    - *What is the average number of rooms?*
  - sum s5aq5, det
    - *What is the median number of rooms?*

Practice 3 – double tabulations

- Please practice the following codes in Stata on your own:
  - tab ur poverty
    - *How many households in rural areas are non-poor?*
  - tab ur poverty, cell
    - *What percent of households **in the sample** live in rural areas and are non-poor?*
  - tab ur poverty, row
    - *What percent of households **who live in rural areas** are non-poor?*
  - tab ur poverty, col
    - *What percent of households **who are non-poor** live in rural areas?*

Practice 4 – using “if”

- Please practice the following codes in Stata on your own:
  - tab poverty if ur==2
    - *What percent of households **who live in rural areas** are non-poor?*
  - tab poverty if s5aq5>3
    - *What percent of households **with more than 3 rooms** are non-poor?*
  - tab poverty if s5aq5<=3
    - *What percent of households **with 3 rooms or less** are non-poor?*

Summary of Day 1

- Today we learned how to:
  - Browse our data (*browse*)
  - Determine whether each variable is *numeric, categorical, or string* (*browse* or *codebook*)
  - Count the number of observations in the dataset (*count*)

Summary of Day 1

- Today we learned how to:
  - Browse our data (*browse*)
  - Determine whether each variable is *numeric, categorical, or string* (*browse* or *codebook*)
  - Count the number of observations in the dataset (*count*)
  - Explore the characteristics of each variable (*codebook*)
  - Find the descriptive statistics of our data (*summarize* and *summarize, detail*)
  - Find the frequencies of single and double variables (*tabulate*)

Summary of Day 1

- Today we learned how to:
  - Browse our data (*browse*)
  - Determine whether each variable is *numeric, categorical, or string* (*browse* or *codebook*)
  - Count the number of observations in the dataset (*count*)
  - Explore the characteristics of each variable (*codebook*)
  - Find the descriptive statistics of our data (*summarize* and *summarize, detail*)
  - Find the frequencies of single and double variables (*tabulate*)
  - Create a histogram (*histogram*)
  - Use “if” to only look at certain observations

Tomorrow's Plan

- Tomorrow we will learn about:
  - Do-files
  - Creating new variables
  - Collapsing data
  - Importing and exporting data to and from Excel

RWANDA

IFPRI

Introduction to Household Data Analysis

Using Stata to Describe, Transform, and Analyze Data

Kigali, Rwanda

March 1-2, 2022

International Food Policy Research Institute (IFPRI)

Contact: Gracie Rosenbach (g.rosenbach@cgiar.org)

Overview of Day 2

- Yesterday we worked with *Household-level* data (aka *wide* data)
  - Information was about the entire household
  - For example, we looked at: What percent of households live in rural areas? How many rooms does the household have?

Overview of Day 2

- Yesterday we worked with *Household-level* data (aka *wide* data)
  - Information was about the entire household
  - For example, we looked at: What percent of households live in rural areas? How many rooms does the household have?

- Today we will work with *Person-level* data (aka *long* data)
  - We will use the household roster that can link back to the household-level data we used yesterday
  - For example, we can look at: What percent of people are female? What is the average age?

Overview of Day 2

- Yesterday we worked with *Household-level* data (aka *wide* data)
  - Information was about the entire household
  - For example, we looked at: What percent of households live in rural areas? How many rooms does the household have?
- Today we will work with *Person-level* data (aka *long* data)
  - We will use the household roster that can link back to the household-level data we used yesterday
  - For example, we can look at: What percent of people are female? What is the average age?
- Will we have more or less observations today?

Day 2 Data

- Let’s open the data to find out!

- Double-click to open file *cs_S1_S2_S3_S4_S6A_S6E_Person.dta* (found on your desktops)

Practice 1 – review of Day 1

- Please practice the following codes in Stata on your own:
  - count
    - *How many observations are in the dataset?*
  - tab s1q1
    - *How many males are in this dataset?*
  - sum s1q3y
    - *What is the average age in this dataset?*

Do-Files

- A do-file contains the Stata commands that you wish to execute

Do-Files

- A do-file contains the Stata commands that you wish to execute

- Executing a do-file is the same as executing a series of commands in Window 2 (Command Window), only you have a permanent record of your commands

Do-Files

- A do-file contains the Stata commands that you wish to execute
- Executing a do-file is the same as executing a series of commands in Window 2 (Command Window), only you have a permanent record of your commands
- This allows you to quickly reproduce work you have already done and go from there, *especially once you start transforming variables*

Do-Files

- A do-file contains the Stata commands that you wish to execute
- Executing a do-file is the same as executing a series of commands in Window 2 (Command Window), only you have a permanent record of your commands
- This allows you to quickly reproduce work you have already done and go from there, *especially once you start transforming variables*
- Double click to open *Rwanda_StataTraining_L1_DescribingData.do*

Practice 2 – recode, replace, and rename

- Please practice the following codes in Stata on your own:
  - recode s1q3y 101 102 103 104 105 106 107 108 109=.
    - *All values greater than 100 are now missing*
  - replace s1q3y=. if s1q3y>100
    - *None are changed, because you already changed them in the 'recode' command above*
  - rename s1q3y age
    - *Now the name of the variable is more intuitive ("age")*
  - label variable age "Age of household member"
    - *Now the label variable is more intuitive*

Practice 3 – creating a new variable

- Please practice the following codes in Stata on your own:
  - gen child=.
  - replace child=1 if age<=15
  - replace child=0 if age>15
  - replace child=. if age==.
  - lab var child “Is the household member 15 years or younger?”
  - lab define no_yes 0 “No” 1 “Yes”
  - lab values child no_yes
  - codebook child

Practice 4 – bar graphs

- Please practice the following codes in Stata on your own:
  - graph bar child, over(province) – *which province has the highest % of children?*
  - graph bar s1q1, over(province) – *the y-axis is wrong!*
  - gen woman=.
  - replace woman=1 if s1q1==2
  - replace woman=0 if s1q1==1
  - lab var woman “Is the household member a woman”
  - lab values woman no_yes
  - graph bar woman, over(province) – *which province has the lowest % of women?*

Practice 5 – collapse

- Please practice the following codes in Stata on your own:
  - gen hhh_female=0
  - replace hhh_female=1 if woman==1 & s1q2==1

  - collapse (count) pid (min) age (mean) child woman (max) hhh_female, by(hhid)

  - ren pid hhsize
  - lab var hhsize “Household size”

  - sum hhsize
    - *What is the average household size?*

Merging

- What if we want to know if the percent of children in a household is associated with a household's poverty status?

- We have to combine two datasets together

Merging

- What if we want to know if the percent of children in a household is associated with a household's poverty status?

- We have to combine two datasets together

cs_S1_S2_S3_S4_S6A_S6E_Person (collapsed; n=14,580)			**cs_S0_S5_Household (n=14,580)**			**merged data (n=14,580)**		
hhid	perc_children		hhid	poverty		hhid	perc_children	poverty
214528	.6		214528	Non-poor		214528	.6	Non-poor

Practice 6 – merge (1)

- Please practice the following codes in Stata on your own:
  - merge 1:1 hhid using “F:\cs_S0_S5_Household.dta“
  - drop _merge
  - bysort poverty: sum perc_children

Practice 7 – save, export, import, encode

- Please practice the following codes in Stata on your own:
  - save "F:\eicv_merged_data.dta", replace
  - export excel using "F:\eicv_merged_data.xlsx", firstrow(variables) replace
  - import excel "F:\eicv_merged_data.xlsx", sheet("Sheet1") firstrow clear
  - encode ur, generate(urban)

Practice 8 – use and merge(2)

- Please practice the following codes in Stata on your own:
  - use "F:\cs_S0_S5_Household.dta", clear

  - merge 1:m hhid using "F:\cs_S1_S2_S3_S4_S6A_S6E_Person.dta“
  - drop _merge

Summary of Day 2

- Today we learned how to:
  - Create a new variable (*generate*) and label it (*label variable*)
  - Replace and recode values (*replace* and *recode*)
  - Create value labels (*label define*) and apply them to a variable (*label values*)

Summary of Day 2

- Today we learned how to:
  - Create a new variable (*generate*) and label it (*label variable*)
  - Replace and recode values (*replace* and *recode*)
  - Create value labels (*label define*) and apply them to a variable (*label values*)
  - Create a bar graph (*graph bar*)
  - Collapse data (*collapse*)
  - Rename a variable (*rename*)

Summary of Day 2

- Today we learned how to:
  - Create a new variable (*generate*) and label it (*label variable*)
  - Replace and recode values (*replace* and *recode*)
  - Create value labels (*label define*) and apply them to a variable (*label values*)
  - Create a bar graph (*graph bar*)
  - Collapse data (*collapse*)
  - Rename a variable (*rename*)
  - Merge data (*merge*)
  - Export to excel (*export excel*) and import from excel (*import excel*)
  - Encode a variable/make it categorical (*encode*)

Tomorrow's Plan

- Tomorrow we will learn about:
  - Outliers
  - Correlations
  - T-tests
  - Regressions
  - Figures (bar and line graphs)

RWANDA

IFPRI

Introduction to Household Data Analysis

Using Stata to Describe, Transform, and Analyze Data

Kigali, Rwanda

March 1-2, 2022

International Food Policy Research Institute (IFPRI)

Contact: Gracie Rosenbach (g.rosenbach@cgiar.org)

Overview of Day 3

- Today we are back to working with *Household-level* data

- We will be using NISR's 'poverty file' in which they calculate:
  - Adult equivalent
  - Consumption
  - Poverty status
  - Food share

Definitions – household consumption

- Household consumption refers to the monetary value of the amount of food and non-food items consumed by the household in a given time period (usually monthly)

Definitions – household consumption

- Household consumption refers to the monetary value of the amount of food and non-food items consumed by the household in a given time period (usually monthly)

- We usually use *household consumption* rather than household income to assess household welfare because income might fluctuate throughout the year, but consumption usually remains somewhat fixed

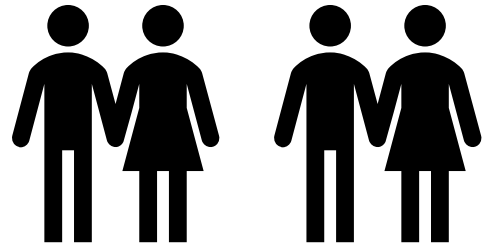
Definitions – adult equivalent

- Adult equivalent is an alternative measure to household size
- Adult equivalent is a better measure for assessing household welfare since each household member has different consumption needs

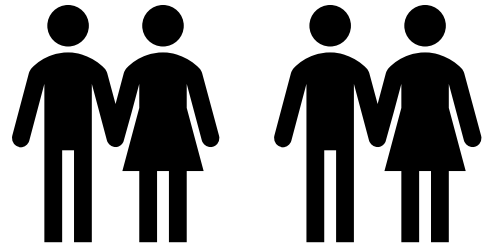
Definitions – adult equivalent

- Adult equivalent is an alternative measure to household size
- Adult equivalent is a better measure for assessing household welfare since each household member has different consumption needs
- It assumes that different household members consume different amounts
  - Adult men = 1
  - Adult women = 0.7
  - Children = 0.5

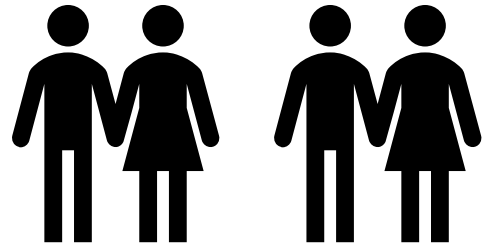
Definitions – adult equivalent

	Household 1	Household 2
		
Monthly consumption	400,000 RWF	400,000 RWF

Definitions – adult equivalent

	Household 1	Household 2
		
Monthly consumption	400,000 RWF	400,000 RWF
Household size	4	4
Consumption per capita	100,000 RWF	100,000 RWF

Definitions – adult equivalent

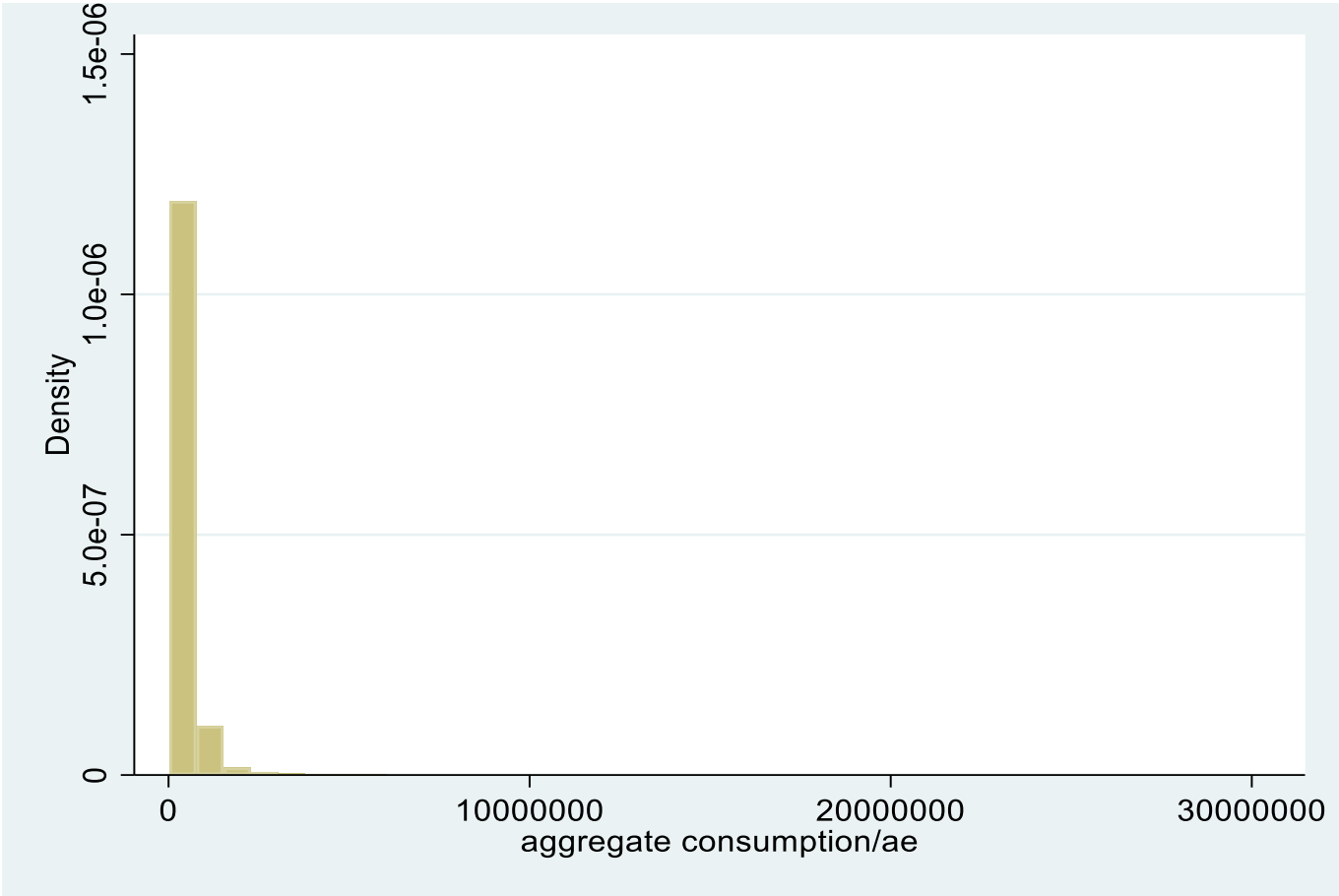
	Household 1	Household 2
		
Monthly consumption	400,000 RWF	400,000 RWF
Household size	4	4
Consumption per capita	100,000 RWF	100,000 RWF
Adult equivalent calculation	Adult male (1) + Adult female (0.7) + Child 1 (0.5) + Child 2 (0.5)	Adult male 1 (1) + Adult female 1 (0.7) + Adult male 2 (1) + Adult female 2 (0.7)
Adult equivalent	2.7	3.4
Consumption per adult equivalent	148,148 RWF	117,647 RWF

Definitions – outliers

- Continuous data (e.g. household consumption) can be messy and contain *outliers*

- An outlier is an observation that is very different from all other observations

Figure: Histogram of monthly consumption per adult equivalent from EICV5 poverty file



Definitions – outliers

- There are 3 main reasons an outlier could exist in the data:
    - Input error (e.g. data collector typed that an egg costs 1000 RWF instead of 100)

Definitions – outliers

- There are 3 main reasons an outlier could exist in the data:
  - Input error (e.g. data collector typed that an egg costs 1000 RWF instead of 100)
  - Confusion about the question by the respondent (e.g. a respondent double counted some of the food his household consumed)

Definitions – outliers

- There are 3 main reasons an outlier could exist in the data:
  - Input error (e.g. data collector typed that an egg costs 1000 RWF instead of 100)
  - Confusion about the question by the respondent (e.g. a respondent double counted some of the food his household consumed)
  - Best guesses (e.g. a respondent doesn't know how much his eggs would sell for at the market since he consumed them all, so he guesses 1000 RWF)

Definitions – outliers

- Cleaning outliers is important because they can skew the data. Cleaning requires you to:

Definitions – outliers

- Cleaning outliers is important because they can skew the data. Cleaning requires you to:

  - Identify outliers – 2 common approaches are:
    - 1% and/or 99% percentile (depending on the skew)
    - 3 standard deviations away from the mean

Definitions – outliers

- Cleaning outliers is important because they can skew the data. Cleaning requires you to:
  - Identify outliers – 2 common approaches are:
    - 1% and/or 99% percentile (depending on the skew)
    - 3 standard deviations away from the mean
  - Decide how to clean them – 3 common approaches are:
    - Change the values to the median
    - Change the values to missing
    - Drop the observations

Day 3 Data

- Let’s open the data to take a look at what the NISR poverty file looks like

- Double-click to open file *EICV5_Poverty_file.dta* (found on your desktops)

Practice 1 – review of Day 1 and Day 2

- Please practice the following codes in Stata on your own:
  - count
    - *How many observations are in the dataset?*
  - sum member
  - sum ae
    - *How do the averages of household size and adult equivalent compare?*
  - sum cons1_ae, det
  - hist cons1_ae
    - *Are there outliers in the consumption variable?*
  - replace cons1_ae=256655.8 if cons1_ae>2600404

Practice 2 – correlations and scatterplots

- Please practice the following codes in Stata on your own:
  - pwcorr cons1_ae member, sig
  - scatter cons1_ae member
  - twoway (scatter cons1_ae member) (lfit cons1_ae member)

Practice 3 – correlations, t-tests, and regressions

- Please practice the following codes in Stata on your own:
  - ttest cons1_ae, by(ur)
  - regress cons1_ae member
  - pwcorr cons1_ae member, sig
    - *The p-value should be the same in this regression and correlation since nothing else is being held constant in the regression!*

Practice 4 – regressions

- Please practice the following codes in Stata on your own:
  - codebook ur
  - recode ur 2=0
  - regress cons1_ae member foodshare1 ur
  - codebook province
  - regress cons1_ae member foodshare1 ur i.province

Summary of Day 3

- Today we learned how to:
  - Measure household welfare using consumption per adult equivalent
  - Clean outliers (*summarize, detail* then *replace*)

Summary of Day 3

- Today we learned how to:
  - Measure household welfare using consumption per adult equivalent
  - Clean outliers (*summarize, detail* then *replace*)
  - Run correlations (*pwcorr, sig*)
  - Run t-tests (*ttest, by*)

Summary of Day 3

- Today we learned how to:
  - Measure household welfare using consumption per adult equivalent
  - Clean outliers (*summarize, detail* then *replace*)
  - Run correlations (*pwcorr, sig*)
  - Run t-tests (*ttest, by*)
  - Run regressions (*regress*), including:
    - What types of variables make sense to be added to a regression
    - Checking whether variables are okay as they are or need to be transformed
    - How to transform the variables so they are suitable for the regression

Thank you!

- Keep in touch with IFPRI Rwanda to hear about future learning events
  - Twitter: @IFPRIRwanda
  - Website: Rwanda.ifpri.info (and sign up for our newsletter at the bottom of the page!)