

# **Basic Data Analysis Course on Stata**

## **Summarization of data in Stata**

Md Monowar Hossain

Institute of Statistical Research and Training (ISRT)  
University of Dhaka  
[mmhossain1@isrt.ac.bd](mailto:mmhossain1@isrt.ac.bd)

January 04, 2025

## Section 1

# Structure of Data

# Cross-section data

- **Multiple individuals** observed at the **same time point**.
- Examples: starting salary of university graduates in the year 2016, GDP per capita for world countries in the year 2000, profits of tech firms last year.

# Time series data

- **A single individual observed at multiple points in time.**
- Examples: Inflation for Bangladesh over the last 10 years (2008-2017), Apple's profit each month in the last year.

# Panel Data

- Basic characteristics:
  - Involves regularly repeated observations on the same individuals.
  - In microeconomics applications, individuals are typically people, households, firms, etc., and repeated observations are different time periods.
  - Examples: Profits for all firms in the S&P 500 observed from 1990-2010, unemployment rate for OECD countries observed over the last 5 years.
  - Two dimensions – cross-section and time series. Typically, **N individuals** are observed at **T regular time periods**. Also known as **longitudinal data** or **repeated measures**.

# Panel Data

- Types of panel data:
  - Panel data can be **balanced**, meaning all individuals are observed in all time periods ( $T_i = T$  for all  $i$ ) or **unbalanced** meaning not all observations are observed in all time periods ( $T_i \neq T$ ).
  - The dataset may be a **short panel** (few time periods and many individuals) or a **long panel** (many time periods and few individuals) or **both** (many individuals and many time periods). This distinction has consequences for both estimation and inference.

## Section 2

### Review

## About the data

- We are going to use the national-scale dataset `wm.dta` from Bangladesh Multiple Indicator Cluster Survey (MICS) 2019.
- The data were collected by the Bangladesh Bureau of Statistics in cooperation with the UNICEF Bangladesh, as part of the global MICS programme.
- The survey employed a two-stage stratified cluster sampling approach where the 64 districts were the sampling strata.
- The 2011 national census enumeration areas (EAs) were defined as the primary sampling units (clusters) and households as the secondary sampling units.
- In the first stage of sampling, 3220 EAs were selected with probability proportional to size (PPS) method.
- In the second stage, a systematic sample of 20 households was obtained from each of the selected EAs, which led to a final sample of 64,400 households for the survey.

# One-way table for qualitative data

```
use wm.dta, clear  
  
tab HH7  
  
tab HH7 [iweight=wmweight]  
/* iweight means 'importance weight' */
```

# One-way table for quantitative data

```
recode WB4 (15/20 = 1 "15-20") (20/25 = 2 "20-25") ///
(25/30 = 3 "25-30") (30/35 = 4 "30-35") ///
(35/40=5 "35-40") (40/45=6 "40-45") ///
(45/50=7 "45-50"), generate(agegroups) label(agegrp)
```

```
tab agegroups [iweight=wmweight]
```

## Two-way table for qualitative data

```
tab HH6 welevel [iweight=wmweight]
```

```
tab HH6 welevel [iweight=wmweight], row
```

```
tab HH6 welevel [iweight=wmweight], col
```

```
tab HH6 welevel [iweight=wmweight], cell
```

# Descriptive statistics

```
summarize WB4
```

```
summarize WB4, detail
```

```
summarize WB4 [aweight=wmweight]
```

```
summarize WB4 [aweight=wmweight], detail
```

# Descriptive statistics

```
by HH6, sort: summarize WB4, detail
```

```
by HH6, sort: summarize WB4 [aweight=wmweight], detail
```

```
codebook HH6
```

```
codebook HH7
```

```
keep if HH6==1
```

```
summarize WB4
```

```
keep if HH6==1 & HH7==10
```

```
summarize WB4
```

## Section 3

**svyset**

# Use Case:

- **fweight**: Count-based summaries.
- **iweight**: Importance scaling.
- **aweight**: Precision-weighted analysis.
- **pweight**: Survey design adjustments.

## Comparison Table of Weight Types

Weight Type	Purpose	Adjusts Vari-ance?		Use Case	Example
		No	Yes		
<b>fweight</b>	Represents frequency of identical cases	No		Aggregated / grouped data	Dataset with summary counts of individuals.
<b>iweight</b>	Reflects importance of observations	No		Scaled exploratory analysis	Market size in company revenue analysis.
<b>aweight</b>	Inverse variance weights for precision	Yes		Regression models with heteroscedasticity	Blood pressure studies with equipment reliability.
<b>pweight</b>	Corrects for unequal sampling	Yes		Survey data analysis	National health surveys with complex

## Declare survey design for dataset: svyset

svyset manages the survey analysis settings of a dataset. You use svyset to designate variables that contain information about the survey design, such as the sampling units and weights.

```
svyset [pw=wmweight], psu(WM1) strata(HH7A)
/* WM1 is the cluster number
HH7A is the district */

tab HH7 [iweight=wmweight]
svy: tab HH7
/* both the commands give the same result */

svy: tab HH6 welevel, row pearson
/* pearson chi-square */
```

# Multiple response analysis

```
use multpleresponse.dta, clear
```

```
ssc install mrtab
```

```
mrtab OTHER2_1 OTHER2_2 OTHER2_3 OTHER2_4 OTHER2_5 ///
OTHER2_6 OTHER2_7 OTHER2_8 OTHER2_9 OTHER2_10 ///
OTHER2_11 OTHER2_12 OTHER2_13 OTHER2_14 OTHER2_15 ///
OTHER2_16 OTHER2_17 OTHER2_18 OTHER2_19 OTHER2_20 ///
OTHER2_21 OTHER2_22 OTHER2_23 OTHER2_24 OTHER2_25 ///
OTHER2_26 OTHER2_27
```

# Exporting results

- Copy table to a excel file
- Copy table to a word file
- Copy figure to a word file
- Create a pdf file of output

## Exporting results

### ① Copy table to Excel or Word:

- In Stata, right-click on the table and select “Copy as table”. Paste it into Excel or Word.

### ② Export table to Word:

Install **asdoc** if not already installed:

```
ssc install asdoc
```

Then use:

```
asdoc tab HH7, replace
```

## About the data

- We will use the national-scale dataset bdhs2022.dta from the **Bangladesh Demographic and Health Survey (BDHS) 2022**.
- The survey was conducted by the **National Institute of Population Research and Training (NIPORT)** in collaboration with **ICF International** and funded by **USAID**.
- **Two-stage stratified cluster sampling** was employed, using the **2011 national census enumeration areas (EAs)** as the primary sampling units (PSUs) and **households** as the secondary sampling units.
- In the first stage, **clusters were selected with the PPS method**, and in the second stage, a **systematic sample of households** was taken. The final sample consists of **roughly 30,000 households**.

# THANK YOU