Training on Statistical Tools for Research: Stata

Survey Data Analysis in Stata

Md Monowar Hossain

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

September 26, 2025

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 companies last year.

Definition:

Multiple individuals observed at the **same point in time**.

Example 1: One variable per individual

Starting salary (in \$1000s) of 5 university graduates in **2016**.

Graduate	Salary (2016)
G1	35
G2	40
G3	32
G4	45
G5	38

Example 2: Multiple variables per individual

Starting salary, GPA, and major of 5 university graduates in **2016**.

Graduate	Salary (2016, $1000s)	GPA	Major
G1	35	3.5	Economics
G2	40	3.8	Statistics
G3	32	3.2	Accounting
G4	45	3.9	Finance
G5	38	3.6	Marketing

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.

Time series data

- **Definition:** A single individual observed at **multiple time points**.

Example: Bangladesh inflation rate (%) from 2010–2017.

Year	Bangladesh Inflation (%)
2010	7.5
2011	8.0
2012	6.8
2013	7.2
2014	6.5
2015	6.0
2016	5.8
2017	5.5

Panel Data

- Basic characteristics:
  - Involves regularly repeated observations on the same individuals.
  - In microeconomics applications, individuals are typically people, households, companies, etc., and repeated observations are different time periods.
  - Examples: Profits for all companies 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.
  - 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.

Panel Data

- **Types of panel data:**

1. Balanced Panel

- Every individual is observed in every time period.
- Example: Suppose we track **profits of 3 companies** (A, B, C) for **3 years** (2018–2020).

companie	2018	2019	2020
A	10	12	15
B	8	9	11
C	20	21	25

2. Unbalanced Panel

- Some individuals are missing observations for certain time periods.
- Example: Same as above, but companie C's 2019 data is missing.

companie	2018	2019	2020
A	10	12	15
B	8	9	11
C	20	—	25

3. **Short Panel** (many individuals, few time periods)

- Example: **1000 households** surveyed on **income** for only **2 years** (2019, 2020).
- Small version:

Household	2019	2020
H1	50	52
H2	70	72
H3	60	63
...	...	...
H1000	60	63

4. **Long Panel** (few individuals, many time periods)

- Example: ** companie's annual revenue** tracked for **20 years** (2000–2019).

companie	2001	2002	...	2020
A	50	52	...	60
B	70	72	...	69
C	60	63	...	67

5. Both (Large N and Large T)

- Many individuals tracked for many years.
- Example: **50 countries' GDP per capita** from **1990–2020**.
- Small version:

Country	1990	1991	1992	...	2020
USA	25k	26k	27k	...	60k
India	2k	2.1k	2.3k	...	8k
Brazil	5k	5.2k	5.4k	...	12k
...	...	...	...	...	...
Bangladesh	2k	5.2k	3.4k	...	12k

Review

About the Data

- Dataset: `wm.dta` from **Bangladesh MICS 2019**
- Collected by **Bangladesh Bureau of Statistics** and **UNICEF Bangladesh**
- Survey design: **two-stage stratified cluster sampling**
  - 64 districts as **strata**
  - **Primary sampling units (PSUs):** enumeration areas (EAs)
  - **Secondary sampling units (SSUs):** households
- Stage 1: 3220 EAs selected using **probability proportional to size (PPS)**
- Stage 2: **20 households per EA** sampled systematically
- Final sample: **64,400 households**

One-way Table – Qualitative Data

```
tab HH7
tab HH7 [iweight=wmweight]
```

- `HH7` is a categorical variable representing household characteristics.
- `[iweight=wmweight]` applies **importance (frequency) weights**, adjusting counts to reflect **population representation**.
- Weighted tables give a **nationally representative frequency distribution**.

One-way Table – 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]
```

- Groups numeric ages into categories for easier interpretation.
- Weighted table reflects **proportion of population in each age group**.

Two-way Table – 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
```

- Cross-tabulation shows the **relationship between two categorical variables**.
- Row, column, and cell percentages provide different perspectives.
- Helps identify **patterns or associations** between variables.

Descriptive Statistics

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

- Default `summarize` gives unweighted statistics.
- `[aweight=wmweight]` applies **analytic weights**, adjusting for population representation.
- `detail` shows **full distribution including percentiles**.
- Weighted descriptive statistics provide **nationally representative means, SDs, and ranges**.

Descriptive Statistics by Group

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

codebook HH6
codebook HH7
```

- Summarizes a quantitative variable **within groups** of `HH6`.
- Useful for comparing subgroups of the population.
- `codebook` provides **metadata and value labels** to understand variables.

svyset

Scenario

- **True population structure of a town:**
  - 80% Poor (800 people if population = 1000)
  - 20% Rich (200 people if population = 1000)
- **True average income:**
  - Poor avg income = **$1000**
  - Rich avg income = **$5000**

1. True Population Average Income

$$\text{Population Mean} = (0.8 \times 1000) + (0.2 \times 5000)$$

$$= 800 + 1000 = 1800$$

True mean income = $1800

2. Sample (Unweighted)

You take **50 poor + 50 rich** (oversampling the rich group).

$$\text{Sample Mean (unweighted)} = \frac{(50 \times 1000) + (50 \times 5000)}{100}$$

$$= \frac{50,000 + 250,000}{100} = \frac{300,000}{100} = 3000$$

This is **biased upward** (gives \$3000 instead of \$1800).

3. Apply Survey Weights (pweight)

- Poor weight $= \frac{80}{50} = 1.6$
- Rich weight $= \frac{20}{50} = 0.4$

Now compute the **weighted mean**:

$$\text{Weighted Mean} = \frac{(1000 \times 50 \times 1.6) + (5000 \times 50 \times 0.4)}{(50 \times 1.6) + (50 \times 0.4)}$$

$$= \frac{(80,000) + (100,000)}{80 + 20} = \frac{180,000}{100} = 1800$$

Weighted average = **$1800 (correct population mean)**

Use Case:

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

Weight Type	Purpose	Variance Adjusted?	Typical Use	Example
fweight	Frequency of identical cases	No	Grouped/ aggregated data	Counts of individuals
iweight	Relative importance of cases	No	Exploratory scaling	Market size in revenue
aweight	Inverse variance (precision)	Yes	Regression (heterosced.)	Blood pressure (measurement)

Weight Type	Purpose	Variance Adjusted?	Typical Use	Example
pweight	Sampling probability correction	Yes	Survey data	National health surveys
Longitudinal	Attrition/ dropout adjustment	Yes	Panel/ longitudinal surveys	Labor market survey (multi-year)

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 multipleresponse.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