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 22, 2025

Survey Analysis in Stata

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.

Survey Design – svyset

```
svyset [pw=wmweight], psu(WM1) strata(HH7A)
```

- `[pw=wmweight]` defines **probability weights**, correcting for **unequal selection probabilities**.
- `psu(WM1)` specifies clusters; `strata(HH7A)` specifies strata.
- Enables **survey-adjusted analysis** for tables, regressions, and tests.

Two-way Table – Survey-adjusted Analysis

```
svy: tab HH6 welevel, row pearson
```

- `svy:` performs **survey-adjusted analysis**, accounting for weights, clustering, and strata.
- Adjusted Pearson chi-square test provides **valid p-values** for complex survey data.

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
```

- Handles **"select all that apply"** questions.
- `mrtab` generates a **combined frequency table** across multiple response variables.

Exporting Results

- Copy tables to **Excel** or **Word**.
- Copy figures to **Word**.
- Save output as a **PDF** for reporting.
- Ensures **reproducibility and ease of sharing**.

Advanced Graphs

- **Density plot:** `kdensity bwt` – smooth distribution curve.
- **Dot plot:** `dotplot bwt` – each observation as a dot.
- **Cumulative distribution (CDF):** `cumul bwt, gen(cumbwt)` → `line cumbwt bwt`
- **Q-Q plot:** `qnorm bwt` – compare distribution to normal.
- **P-P plot:** `pnorm bwt` – alternative normality check.
- **Scatter matrix:** `graph matrix bwt age lwt` – pairwise scatter plots.
- **Violin plot:** `vioplot bwt, over(smoke)` – density + boxplot combined.
- **Combined plots:** `twoway (scatter ...) (lfitci ...)`
- scatter + regression + CI.

THANK YOU