Lectures 04, 05, 06: Sample weights

Ernesto F. L. Amaral

September 12–19, 2017
Advanced Methods of Social Research (SOCI 420)

Sources: Stata Help & General Social Survey Codebook.

TEXAS A&M UNIVERSITY

Using sample weights

Name	Number of observations collected in the survey	Weight to expand to population size	Weight to maintain sample size
José	1	4	0.8
Maria	1	6	1.2
Total	2	10	2

Sample weight =
Population weight * (Sum of sample weights / Sum of population weights)

Frequency weight in Stata

- **FWEIGHT**
  - Expands survey size to the population size
  - Indicates the number of duplicated observations
  - Used on tables to generate frequencies
  - Can be used in frequency distributions only when weight variable is discrete (no fractional numbers)

```
tab x [fweight = weight]
```

”Importance” weight in Stata

- **IWEIGHT**
  - Indicates the ”importance” of the observation in some vague sense
  - Has no formal statistical definition
  - Any command that supports iweights will define exactly how they are treated
  - Is intended for use by programmers who want to produce a certain computation
  - Can be used in frequency distributions even when weight variable is continuous (fractional numbers)

```
tab x [iweight = weight]
```

Analytic weight in Stata

- **AWEIGHT**
  - Inversely proportional to the variance of an observation
  - Variance of the *j*th observation is assumed to be σ^2/w_j, where w_j are the weights
  - For most Stata commands, the recorded scale of aweights is irrelevant
  - Stata internally rescales frequencies, so sum of weights equals sample size

```
tab x [aweight = weight]
```

More about analytic weight

- Observations represent averages and weights are the number of elements that gave rise to the average

group	x	y	n
1	3.5	26.0	2
2	5.0	20.0	3

- Instead of

group	x	y
1	3	22
1	4	30
2	8	25
2	2	19
2	5	16

- Usually, survey data is collected from individuals and households (not as averages)
  - Thus, aweights are not appropriate for most cases

Sampling weight in Stata

- **PWEIGHT**
  - Denote the inverse of the probability that the observation is included due to the sampling design
  - Indicated for statistical regressions
  - Variances, standard errors, and confidence intervals are estimated with a more precise procedure
    - Estimates robust standard errors

```
regress y x1 x2 [pweight = weight]
```

Summary of Stata weights

WEIGHTS IN FREQUENCY DISTRIBUTIONS		
Weight unit of measurement	**Expand to population size**	**Maintain sample size**
Discrete	**fweight**	**aweight**
Continuous	**iweight**	

WEIGHTS IN STATISTICAL REGRESSIONS should maintain sample size	
Robust standard error	**Adjusted R^2, TSS, ESS, RSS**
pweight	**aweight**
reg y x, robust	**outreg2**

General Social Survey

- GSS targets the adult population (18+) living in households in the United States
- Due to the adoption of the sub-sampling design of non-respondents, a weight must be employed when using the GSS 2004 and after
- There are three continuous weight variables
  - WTSS
  - WTSSNR
  - WTSSALL
- They all maintain the original sample size, even in frequency distributions with ”iweight”

WTSS

- WTSS variable takes into consideration
  - Sub-sampling of non-respondents
  - Number of adults in the household

- In years prior to 2004, a one is assigned to all cases, so they are effectively unweighted
  - Number of adults can be utilized to make this adjustment for years prior to 2004

WTSSNR

- WTSSNR variable takes into consideration
  - Sub-sampling of non-respondents
  - Number of adults in the household
  - Differential non-response across areas
- In years prior to 2004, a one is assigned to all cases, so they are effectively unweighted
  - Number of adults can be utilized to make this adjustment for years prior to 2004
  - No area non-response adjustment is possible

WTSSALL

- WTSSALL takes WTSS and applies an adult weight to years before 2004
- The weight value of WTSSALL is the same as WTSS for 2004 and after
- Researchers who use the GSS data before or after 2004 may consider using the WTSSALL weight variable

```
tab x [aweight = wtssall]
sum x [aweight = wtssall]
```

GSS has a cluster sample

(https://gssdataexplorer.norc.org/pages/show?page=gss%2Fstandard_error)

- First- and second-stage units are selected with probabilities proportional to size
  - Size is defined by number of housing units
- Third-stage units (housing units) are selected to be an equal-probability sample
  - This results in roughly the same number of housing units selected per second-stage sampling unit
- To calculate standard errors correctly, sample design variables must be used
  - Without design variables, Stata will assume a simple random sample and underestimate standard errors

Complex survey design in GSS

(https://gssdataexplorer.norc.org/pages/show?page=gss%2Fstandard_error)

- There are two design variables
  - VSTRAT
  - VPSU
- First-stage unit
  - VSTRAT: Variance Stratum
  - National Frame Areas (NFAs): one or more counties
- Second-stage unit
  - VPSU: Variance Primary Sampling Unit
  - Segments: block, group of blocks, or census tract

Complex survey design in Stata

(https://gssdataexplorer.norc.org/pages/show?page=gss%2Fstandard_error)

- Declare survey design for dataset

```
svyset [weight=wtssall], strata(vstrat) psu(vpsu) singleunit(scaled)
```

- After ”svyset,” you should indicate survey design with the option ”svy” for each table and regression

```
svy: tab x
svy: reg y x1 x2
```

Strata with single sampling unit

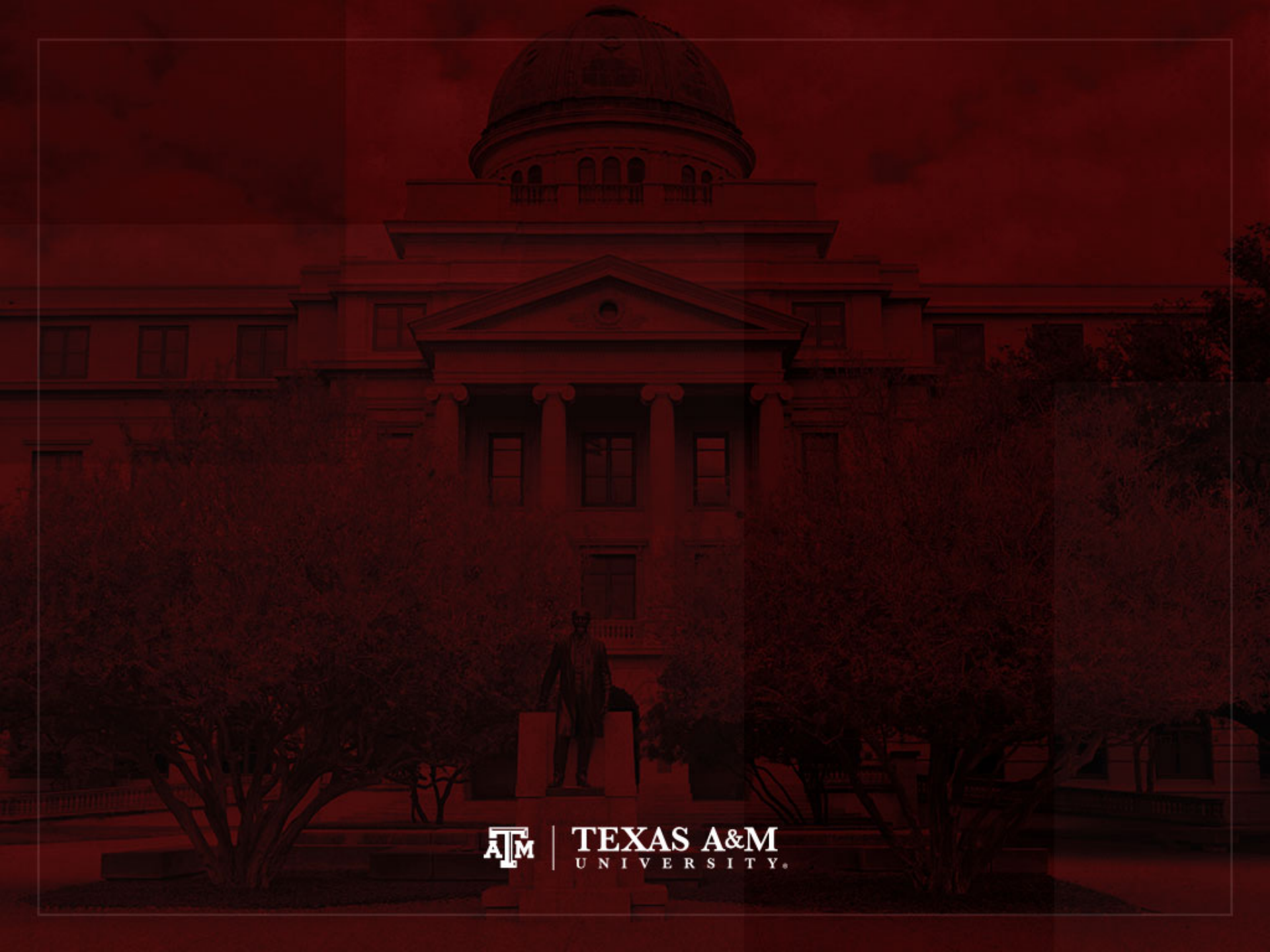
(https://gssdataexplorer.norc.org/pages/show?page=gss%2Fstandard_error)

- VSTRAT and VPSU were created with a minimum of three respondents within a cell
  - If all cases are missing on a variable, you get an error message in Stata
  - ”Missing standard error because of stratum with single sampling unit”
- It is recommended to utilize the ”subpop” option for any subdomain analyses (e.g., for males)

```
svy, subpop(if sex==1): tab x
```

- You can also specify that strata with one sampling unit are ”centered” at grand mean instead of stratum mean

```
svyset [weight=wtssall], strata(vstrat) psu(vpsu) singleunit(centered)
```


TEXAS A&M
UNIVERSITY