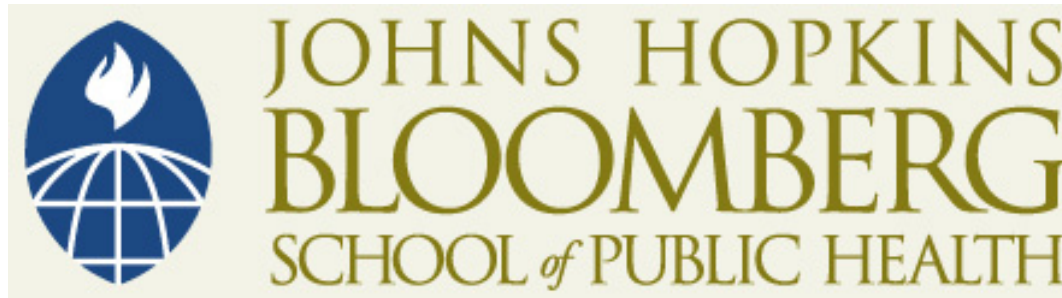


This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike License](https://creativecommons.org/licenses/by-nc-sa/4.0/). Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.



Copyright 2009, The Johns Hopkins University and Saifuddin Ahmed. All rights reserved. Use of these materials permitted only in accordance with license rights granted. Materials provided "AS IS"; no representations or warranties provided. User assumes all responsibility for use, and all liability related thereto, and must independently review all materials for accuracy and efficacy. May contain materials owned by others. User is responsible for obtaining permissions for use from third parties as needed.

Methods in Survey Sampling

Biostat 140.640

Survey Errors

Saifuddin Ahmed, PhD
Dept. of Biostatistics

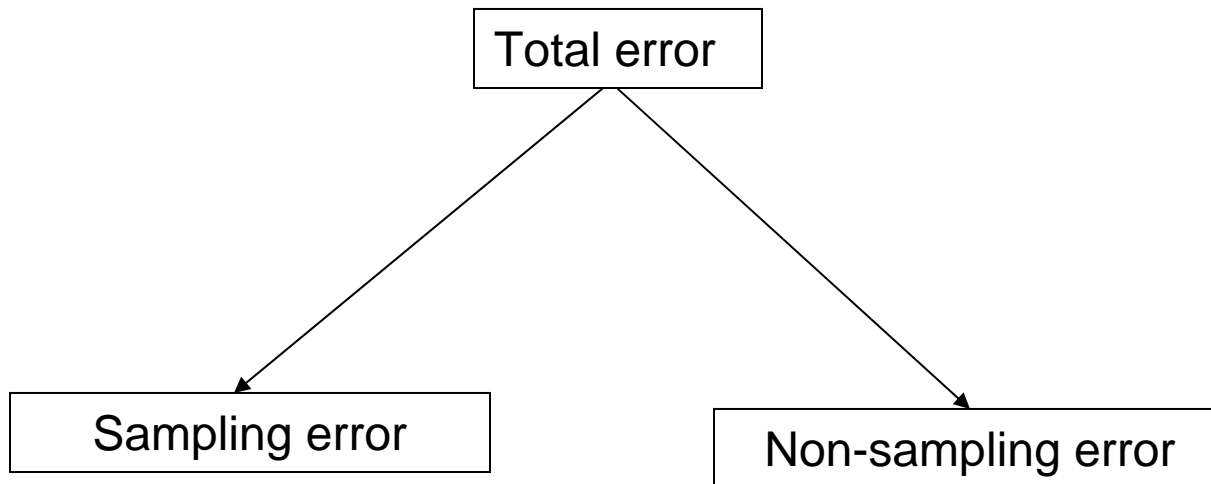
“The information obtained may be incorrect; the definitions and standards used may be loose, unsuitable or wrongly conceived; the households actually visited may not contain a fair sample of the whole population;....”

Bowley and Burnett-Hurst (1915)

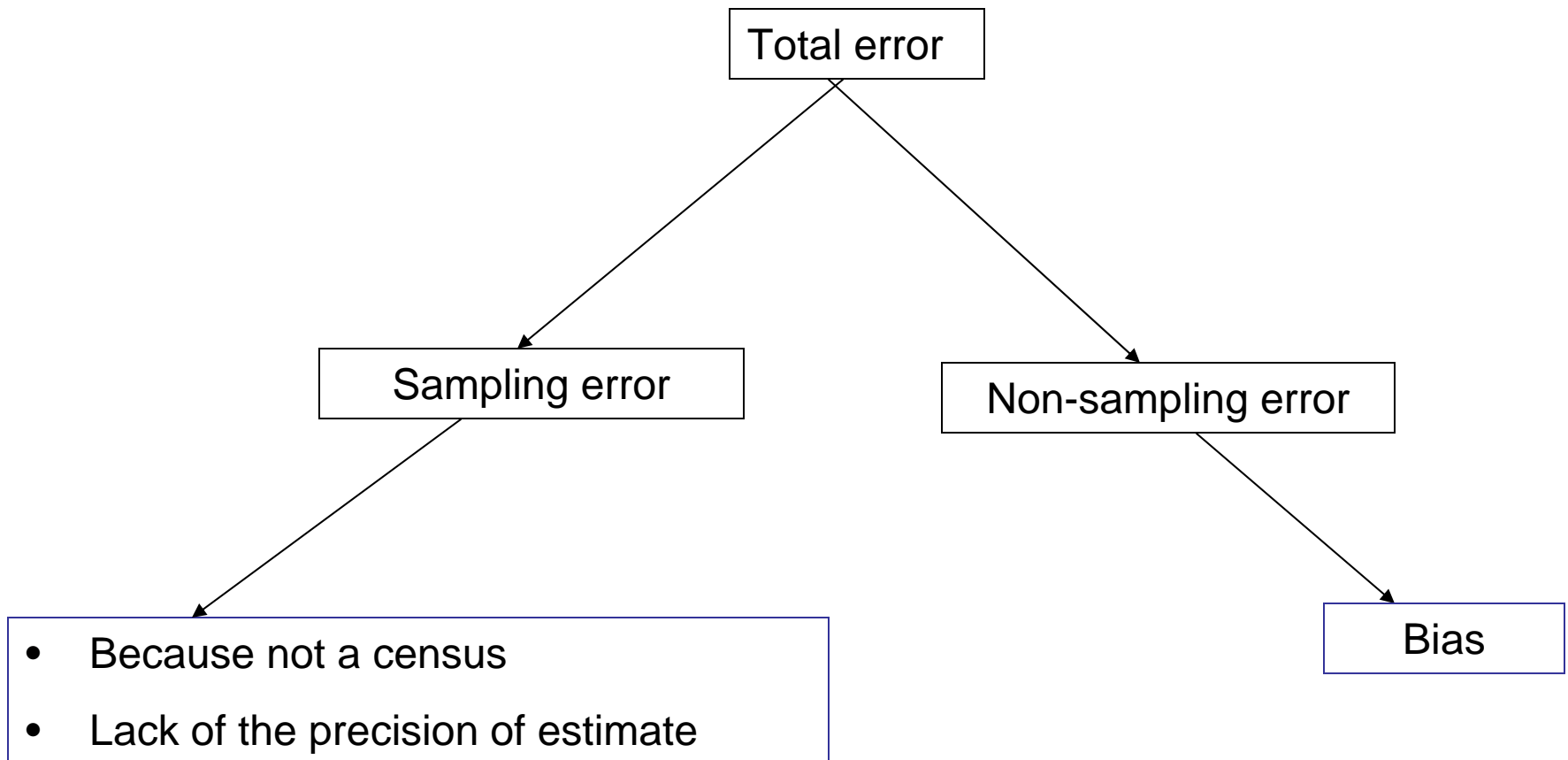
“ For what profiteth a statistician to design a beautiful sample when a questionnaire will not elicit the information desired, or if the universe has not been satisfactorily defined, or the field-force is so badly organized that the results will not be worth tabulating. “

-Deming (1950).

The total error is the sum of all errors about a sample estimate, both sampling and non-sampling.



The total error is the sum of all errors about a sample estimate, both sampling and non-sampling.



Sampling Errors

- The errors are due to conscious choice to study a subset rather than the whole population.
- Depends on the design of the survey.
 - (contrast: stratified vs cluster sampling)
- Depends on sample size.
- **Sampling error** measures the degree to which the sample estimates differ from the expected values. $V(Y) = E[(Y - E(Y))]^2$
- The **square** of **sampling error** is the sampling variance.

Nonsampling Errors

- Any errors possible in the surveys except those by sampling errors.
- The errors are due to mistakes and deficiencies in implementing surveys.

Sources of survey errors

- Survey planning
- Budget consideration
- Sampling
- Preliminary activities (e.g., collection of listing frame)
- Sample design
- Sample selection
- Development of questionnaire and forms
- Instructions for data collection
- Pilot study
- Data collection

- Editing and coding
- Data entry
- Computer data editing and coding/recoding
- Analysis
- Interpretation
- Computation of sample weights
- Report

Measurement of Errors

- **Bias**

- Bias of an estimate y of a population parameter Y is defined as the difference between the mean $E(y)$ of the sampling distribution of y and the true value of the unknown parameter Y . Mathematically,
- $B(y) = E(y) - Y$
- The objective of any survey is to get unbiased estimate of y , i.e. $B(y)=0$.

Measurement of Errors

- **Mean Square Error (MSE)**

- When bias is associated with study, the survey errors are expressed in terms of MSE.
- $MSE = \text{variance} + \text{Bias}^2$
- MSE shows the variability of the survey mean \bar{y} around the true value Y .
- **$RMSE = \sqrt{MSE} = \sqrt{\text{Sampling error}^2 + \text{Bias}^2}$**
- **The aim of any survey is to maintain the greatest possible accuracy by controlling the total error (low RMSE), not just the sampling error.**

Let's look at some of non-sampling errors

Non-response Error

- Unit nonresponse
- Item nonresponse

Unit non-response: The entire set of data is lost: total nonresponse

- a. No eligible respondent at home
- b. Temporarily absent
- c. Refused
- d. Vacant
- e. Not a housing unit
- f. Temporary home/vacation home

Item non-response: Selected items are missing in an otherwise completed questionnaire.

Unit missing (non-response)

```
. *Interview result
```

```
.
```

```
. tab wm7, mis
```

result of women 's interview	Freq.	Percent	Cum.
completed	5,889	94.38	94.38
not at home	263	4.21	98.59
refused	13	0.21	98.80
partly completed	3	0.05	98.85
incapacitated	16	0.26	99.10
other	56	0.90	100.00
Total	6,240	100.00	

Item missing (non-response)

```
. tab age, mis
```

age	Freq.	Percent	Cum.
0	735	2.79	2.79
...//skipped			
...			
97+	3	0.01	99.69
.	81	0.31	100.00
Total	26,329	100.00	

Good [?]



```
. ta agemnthmis [aw=wmweight]   ///Month of birth missing
```

agemnthmis	Freq.	Percent	Cum.
0	3,246	55.12	55.12
1	2,643	44.88	100.00
Total	5,889	100.00	

```
. ta ageyrmis [aw=wmweight]   ///Year of birth missing
```

ageyrmis	Freq.	Percent	Cum.
0	5,649	95.92	95.92
1	240	4.08	100.00
Total	5,889	100.00	

Differentials in missing responses [Education]

```
-> ta melevel agemnthmis [aw=wmweight], row mis
```

education	age month missing		Total
	No	Yes	
	0	1	
none	14.05	85.95	100.00
primary	37.84	62.16	100.00
middle/jss	75.61	24.39	100.00
secondary+	95.52	4.48	100.00
other/missing/dk	50.11	49.89	100.00
Total	55.12	44.88	100.00

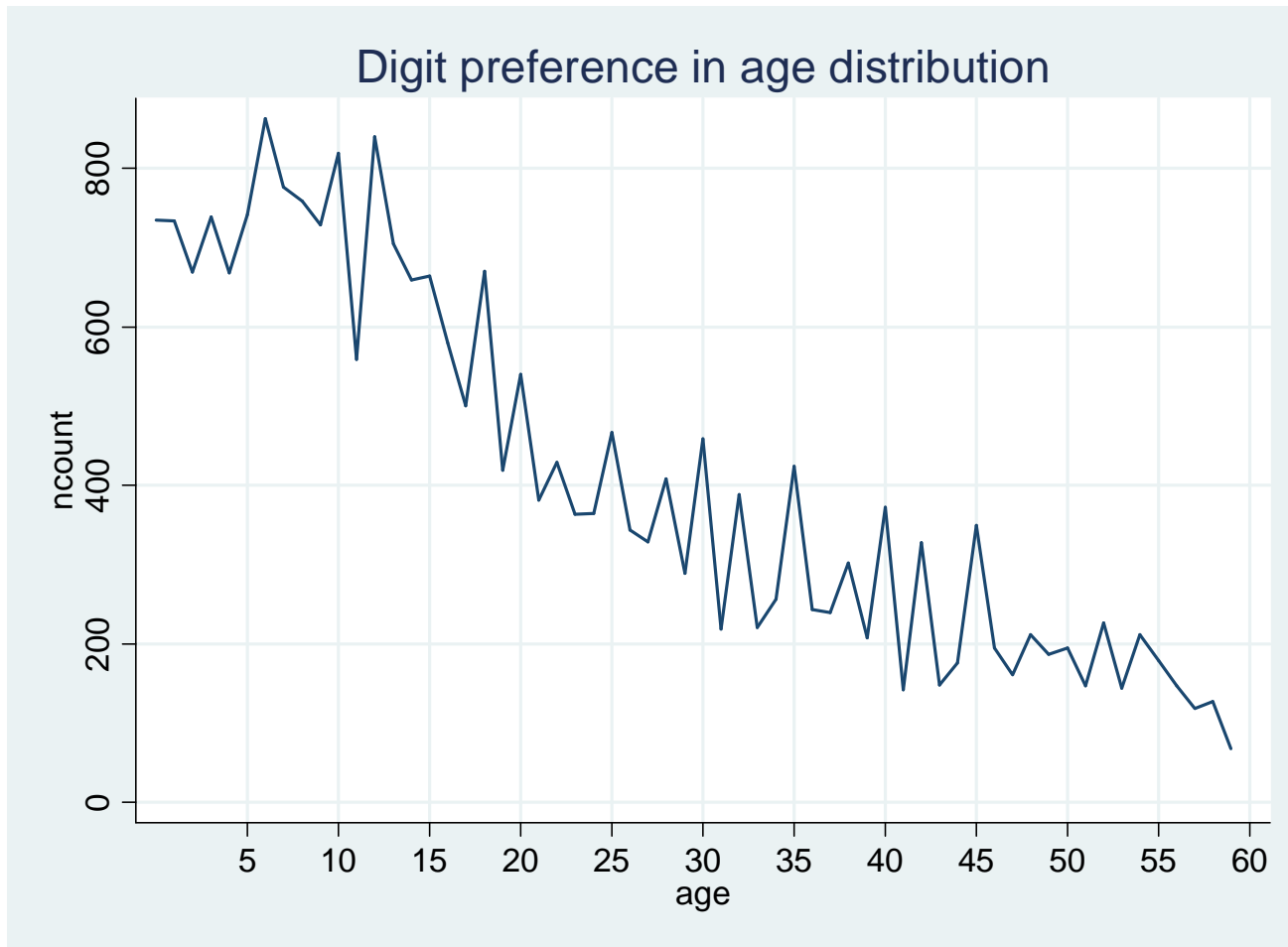
Compare



Differentials in missing responses [Urban/Rural Residence]

area	age month missing		Total
	No 0	Yes 1	
urban	71.47	28.53	100.00
rural	40.55	59.45	100.00
Total	3,245.851	2,643.149	5,889
	55.12	44.88	100.00

Digit preference in age reporting (age heaping at 5 and 0 digits)



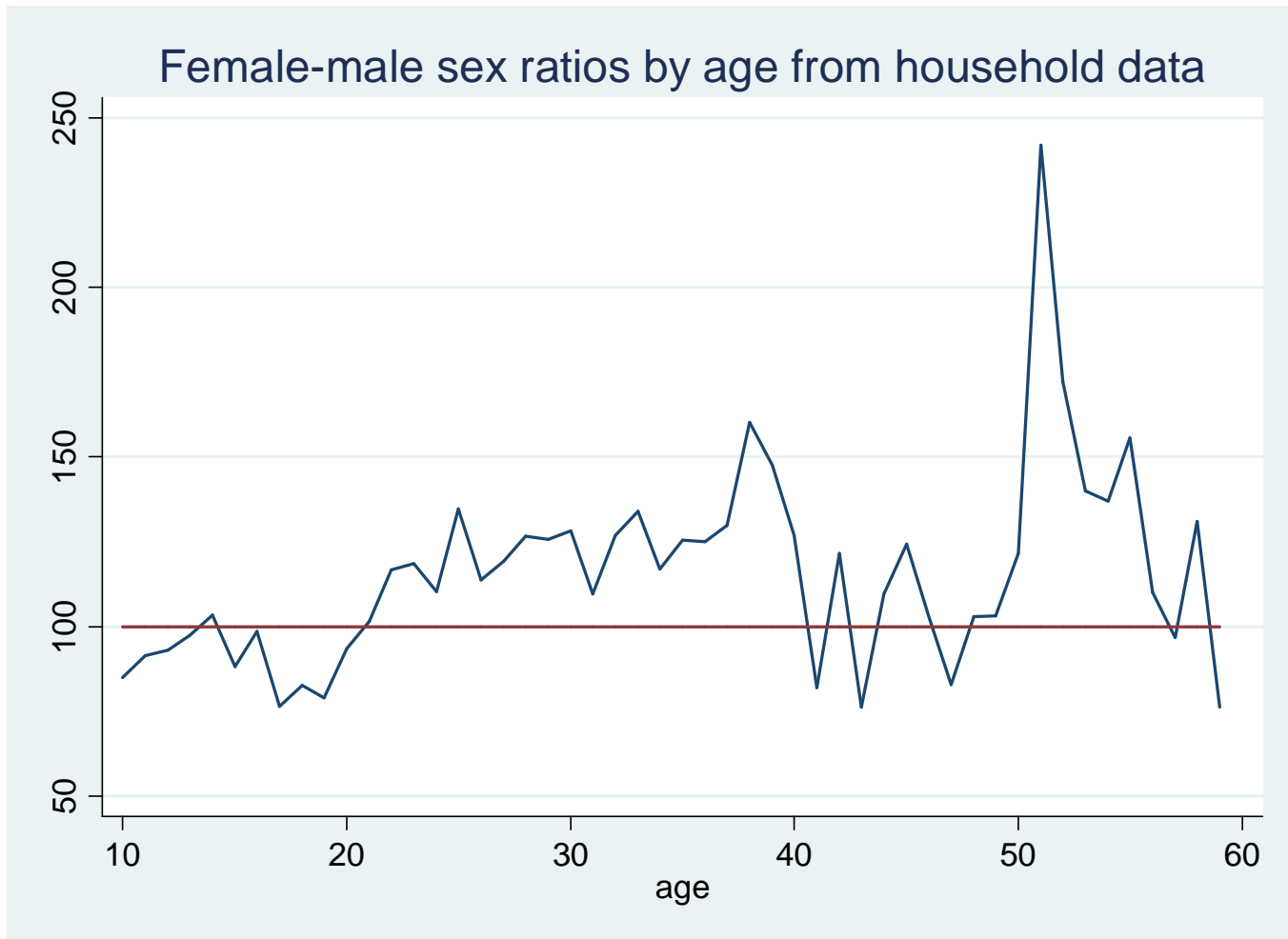
Measuring digit preference

- Whipple's index $\frac{\sum(P_{15} + P_{20} + P_{25} + \dots + P_{40} + P_{45})}{1/5 \sum(P_{15} + P_{16} + P_{17} + \dots + P_{48} + P_{49})} \times 100$
- Whipple's index : 122.10

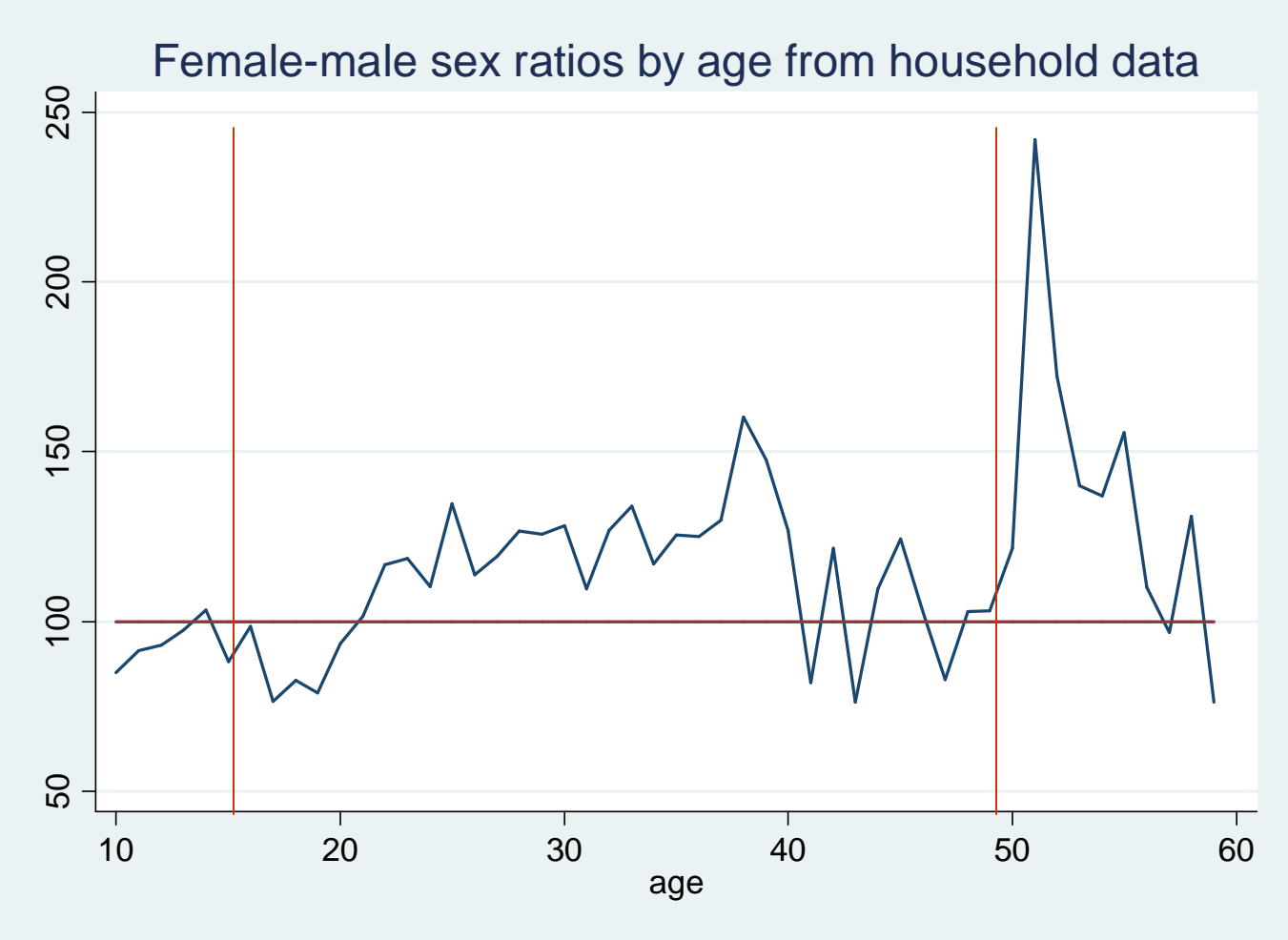
Age displacement

- To reduce the number of interviews, the interviewers often displace the age of the eligible samples so that they become ineligible.
- How to measure?
 - Sex ratios
 - Age ratios

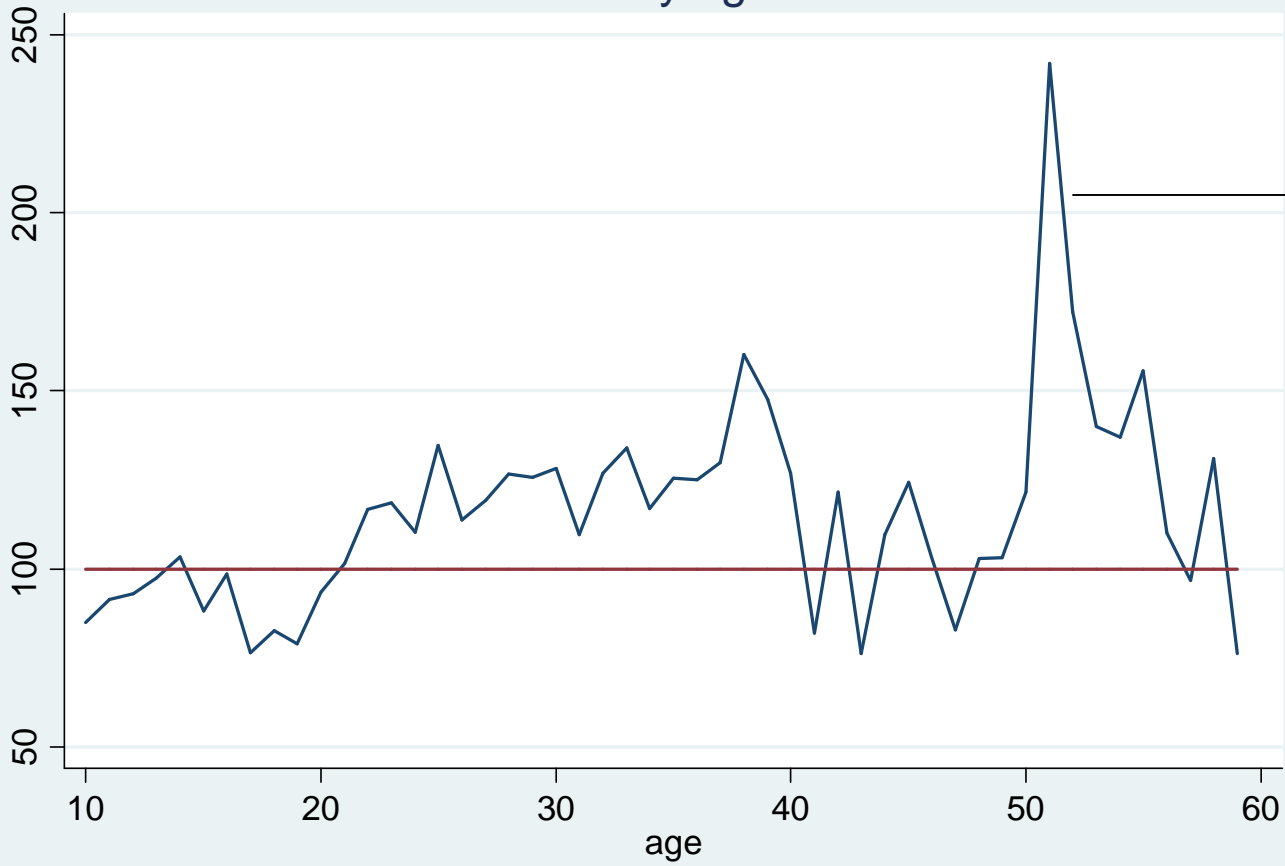
Survey of “reproductive aged women”



The proportion of female respondents were lower in the two extreme end of the reproductive life period: the sign of “age displacement” is evident

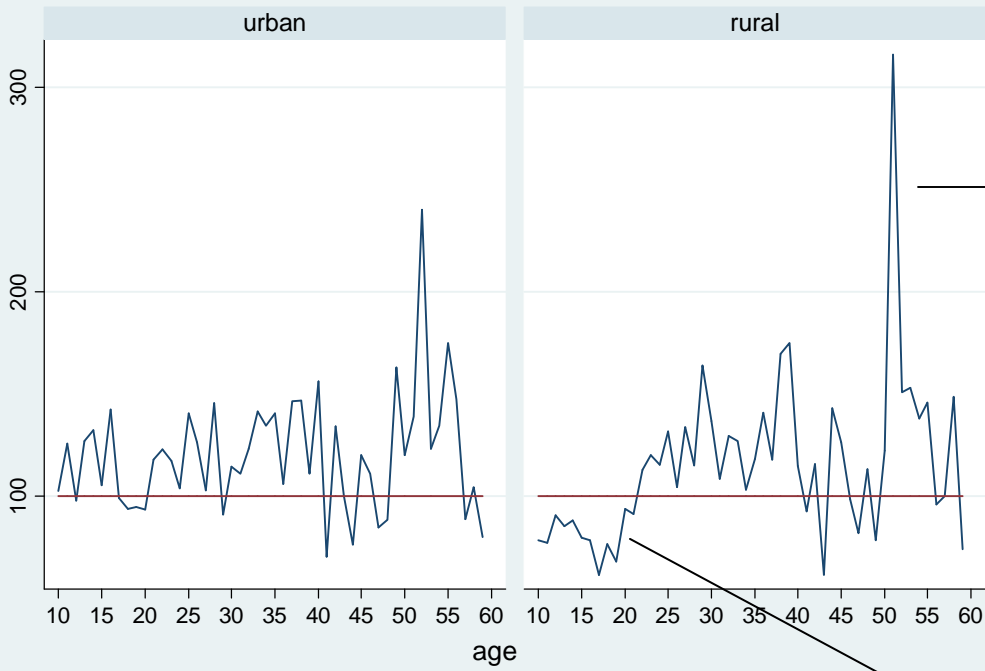


Female-male sex ratios by age from household data



Why this peak?

Female-male sex ratios by age from household data



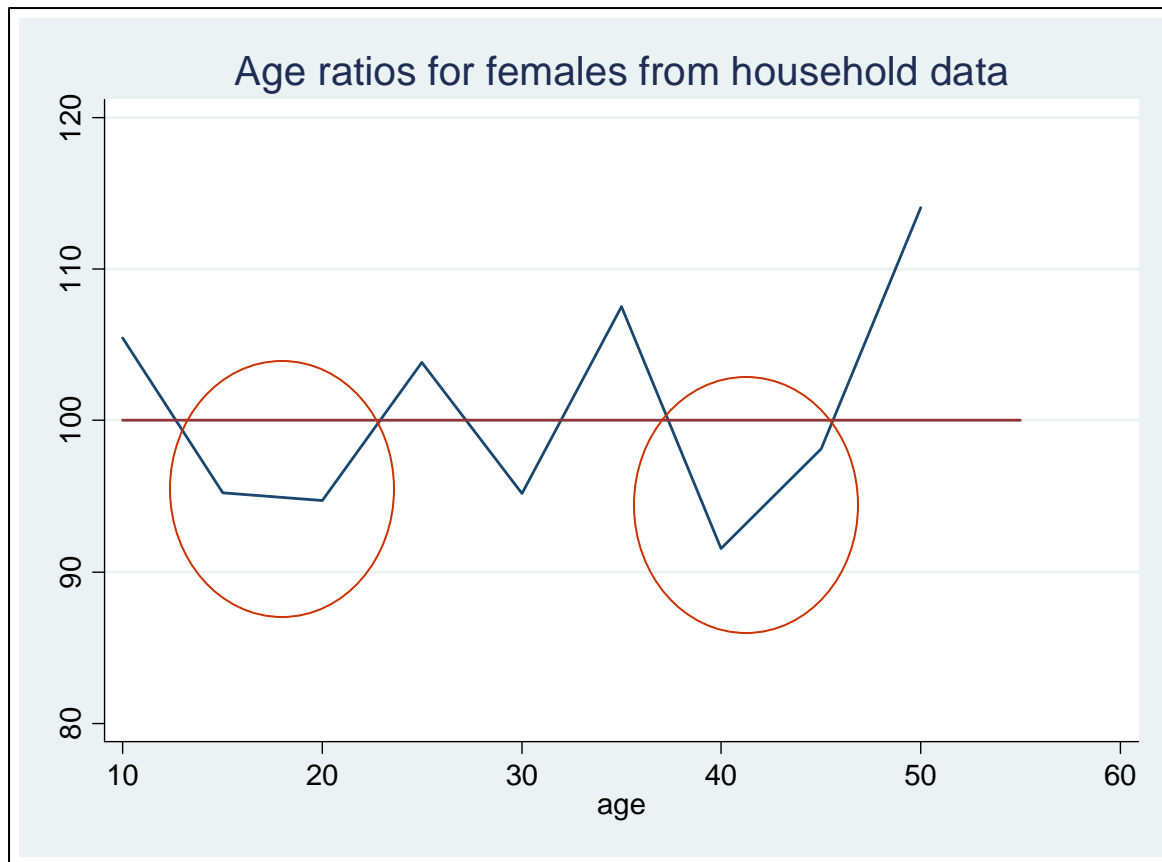
Over-reporting of age(50+)

Under-reporting of age(15-20)

$$\text{Age ratio: } \frac{{}_5P_a}{1/3({}_5P_{a-5} + {}_5P_a + {}_5P_{a+5})} \times 100$$

(e.g., age ratio for age group: 25-29)

$$= P(25-29) / [(1/3)[P(20-24) + P(25-29) + P(30-34)]]$$



```
. list agegr cntfemale ratio_age diff
```

	agegr	cntfem~e	ratio_~e	diff
1.	5-9	1889	.	.
2.	10-14	1731	105.463	-5.463036
3.	15-19	1304	95.25201	4.747993
4.	20-24	1072	94.72754	5.272461
5.	25-29	1019	103.8383	-3.838318
6.	30-34	853	95.20089	4.79911
7.	35-39	816	107.5099	-7.50988
8.	40-44	608	91.56626	8.433739
9.	45-49	568	98.15668	1.843315
10.	50-54	560	114.053	-14.05296
11.	55-59	345	.	.

Age accuracy index = $\text{sum}[\text{abs}(\text{diff})] / \text{no. of age category [non-missing]}$

$$= 55.96 / 9 = 6.2178675$$

Discrepancies in reporting

household has mosquito nets	child slept under bednet last night			Total
	yes	no	dk	
yes	1,081.389 71.54	428.57867 28.35	1.6296357 0.11	1,511.597 100.00
no	49.552587 2.53	1,901.541 97.20	5.3095278 0.27	1,956.403 100.00
Total	1,130.941 32.61	2,330.119 67.19	6.9391635 0.20	3,468 100.00

Household has no mosquito net, but slept under bednet

In summary:

- There are possibilities of errors at every stage of survey implementation.
- You need to evaluate survey data:
 - Extent of missing values (unit and item - both)
 - Any systematic pattern of error
 - Age mis-reporting, heaping
 - Age displacement
 - a significant concern [for infant mortality study]
 - Risk of non-representative sample estimate (biased)
 - Check data for “range-check” (bounded values: e.g., 0,1, 9), skip rule [if q.202=1, q.205=not applicable], logical values, logical patterns