

# Data validation and diagnostics

LECTURE 11

1

---

---

---

---

---

---

---

## Preamble

- Each step of a **survey** can generate **errors** in the data or the published statistics.
- It has been estimated that national statistical institutes spend some 40% of their resources on identifying errors and fixing them.
- This lecture focuses on **errors**, **missing values** (item nonresponse, or a respondent may give a wrong answer), and **unit nonresponse** (a respondent does not answer at all).
- They are largely responsible for overall **data quality**.

2

---

---

---

---

---

---

---

## A bird's-eye view of the survey process

today's focus is on steps 3 and 4

1. **Setting survey objectives**  
(lecture 1)
2. **Questionnaire design and sampling design**
3. **Data collection and data entry**  
the sample is drawn, data are collected from the sampled units and entered into the computer system at the statistical office
4. **Data processing and data analysis**  
collected data are edited, missing and erroneous data are imputed, raising weights are determined
5. **Publication and data dissemination**  
(lecture 15)

3

---

---

---

---

---

---

---

## Types of data errors

4

---

---

---

---

---

---

---

---

## Types of errors

Errors can be classified in several ways:

1. **random** vs. **systematic** errors
2. **influential** vs. and **noninfluential** errors
3. **outliers** vs. **nonoutliers**  
but outliers are not necessarily errors: more on this in lecture 12

5

---

---

---

---

---

---

---

---

## Random vs. systematic errors

- **Random errors** are not caused by a systematic deficiency, but by **accident**.
- In general statistics, the **expectation** of a random error is typically **zero**: errors compensate one another, on average.
- A **systematic error** is an error that occurs frequently between respondents.
- Classical example: **measurement unit error** – e.g. reporting consumption in **kg** instead of the requested **grams**.

6

---

---

---

---

---

---

---

---

### Influential errors

- Errors that have a substantial influence on the statistics of interest are called **influential** errors.
- An influential value is often also an **outlier** (and vice versa) more on this in lecture 12

7

---

---

---

---

---

---

---

---

### Fatal errors

- There is no definition of **'fatal error'** in **statistics**
- We borrow the term 'fatal error' from **computer science**: if you get a fatal error, you generally cannot recover from it, because the computer encounters a problem it cannot resolve.
- So, what could be a 'fatal error' in our context, when we begin the survey on day one, with the data collection on the field?

8

---

---

---

---

---

---

---

---

### Interviewer falsification of survey data

- A **subtle** form of falsification may consist in surveying a **wrong** household member, or in conducting the survey by telephone when face-to-face interviews are required.
- A **severe** form of falsifying is the **fabrication** of entire interviews without ever contacting the respective household.
- Fabricated interviews can have serious consequences for statistics based on the survey data.

9

---

---

---

---

---

---

---

---

## Growing literature

JOURNAL OF DEVELOPMENTAL ENTREPRENEURSHIP • VOL. 7, No. 3 • OCTOBER, 2002

### Interviewer Cheating: Implications for Research on Entrepreneurship in Africa

David E. Harrison and Stefanie I. Krauss

#### Abstract

*Interviewer cheating has seldom been studied or discussed as a problem in the literature. This article therefore begins with a brief review of this problem area, which is of utmost importance especially for entrepreneurship research in a Third World context. In the Third World, the allocation of financial support is often based on interview surveys. After*

10

---

---

---

---

---

---

---

---

## Prevention

3 layers

1. **Study design**
  - Ensuring reasonable interview length
  - Creating positive work conditions and avoiding unrealistic production quotas
  - Interviewer compensations on a per hour, not a per interview basis
2. **Interviewer selection**
  - Employment of interviewers with personal interest in the data quality (e.g., students)
  - Little interviewer experience to reduce likelihood of knowledge on cheating opportunities in the system
3. **Subsequent interviewer inspection**
  - Interview verification via random checks

11

---

---

---

---

---

---

---

---

## Ruling out horror stories

Judge and Schechter (2009)

### Detecting Problems in Survey Data Using Benford's Law

George Judge  
Laura Schechter

#### ABSTRACT

*"It is 15:00 in Nairobi. Do you know where your enumerators are?"*

- "Horror stories are common in which somebody discovers that one (or more) enumerator answers the survey himself rather than actually interviewing households (...)"
- It would be useful to become aware of this as early as possible in the data collection process.

12

---

---

---

---

---

---

---

---

### Benford's law – I/III

- How to recognize survey data irregularities, manipulated outcomes, and abnormal digit and number occurrences?
- The use of [Benford's law](#) offers one possibility.
- Frank Benford was physicist who noticed that the pages in logarithmic tables containing the logarithms of low numbers (1 and 2) were more used than pages containing logarithms of higher numbers (8 and 9).

13

---

---

---

---

---

---

---

---

### Benford's law – II/III

first significant digit

- Benford began to investigate the [distribution of leading digits](#) in a wide range of collections of numbers (numbers on the first page of a newspaper, street addresses, molecular weights...). Here is the law that regulated them:

$$Prob(\text{leading digit} = d) = \log_{10} \left( 1 + \frac{1}{d} \right)$$

- The 'd' in the denominator is responsible of the fact that numbers with a first digit of 1 are observed more often than those starting with 2, 3, ..., 9. Next slide illustrates.

14

---

---

---

---

---

---

---

---

### Benford's law – III/III

first significant digit

- According to B's "law", the leading digit in a number is more likely to be small than large
- If numbers were equally likely to appear as the leading digit, each would have a  $1/9 = 11.1\%$  probability
- According to the law, the **number 1** appears as the leading significant digit about **30%** of the time.
- The **number 9**, less than **5%** of the time.

Dig.	First Digit.
0	. . . .
1	. . . . 0.3010
2	. . . . 0.1761
3	. . . . 0.1249
4	. . . . 0.0969
5	. . . . 0.0792
6	. . . . 0.0669
7	. . . . 0.0580
8	. . . . 0.0512
9	. . . . 0.0458

15

---

---

---

---

---

---

---

---

## Check the data and test hypothesis

- The basic idea of using Benford's law to detect fabricated data is that falsifiers are unlikely to know the law or to be able to fabricate data in line with it.
- A **strong deviation of the leading digits** from Benford's distribution in a dataset indicates that the data might be faked.
- **Hypothesis testing** is what we need.

---

---

---

---

---

---

---

---

Selected surveys exhibit un usually large deviations from the expected 'theoretical' distribution: this is not a good sign.

**Table 6**  
Correlations ( $r$ ), the  $m$  Statistic, Distance  $d'$  and  $a'$ ,  $\chi^2$  Tests, and Kuiper  $V_n$  Tests between Benford's Law and Crop Quantities Produced

Surveys	Observations	$r$	$m$	$d'$	$a'$	$\chi^2$	$V_n$
Bangladesh (1996)	4,522	7.357	0.83	0.065	0.085	0.039	440.92**
China (2002)	294	7.99	0.99	0.016	0.020	0.010	8.45
Mexico (Oct 91)	24,067	10.028	0.84	0.154	0.109	0.059	3059.21**
Mexico (May 92)	22,328	9.659	0.88	0.138	0.135	0.075	3145.66**
Pakistan (Feb 97)	927	874	0.97	0.040**	0.040**	0.036	31.62**
Pakistan (Oct 86)	909	181	0.86	0.100	0.128	0.016	23.21**
Pakistan (Jan 87)	882	563	0.99	0.047	0.062	0.046	20.71**
Pakistan (Mar 87)	854	46	0.94	0.051	0.090	0.028	4.79
Pakistan (Jan 88)	813	728	0.96	0.046	0.069	0.061	41.45**
Pakistan (Mar 88)	809	69	0.92	0.048	0.086	0.020	8.11
Pakistan (Aug 88)	804	559	0.99	0.022	0.038	0.027	15.44
Pakistan (Mar 89)	802	706	0.99	0.035	0.042	0.054	26.23**
Pakistan (Mar 91)	766	200	0.78	0.108	0.141	0.037	29.67**
Poland (1994)	759	691	0.98	0.025	0.044	0.020	22.64**
Pakistan (Oct 91)	726	1,095	0.99	0.022	0.033	0.017	18.72**
Paraguay (1999)	298	1,412	0.99	0.026	0.036	0.025	38.56**
Paraguay (2002)	223	1,632	0.97	0.050	0.065	0.051	101.34**
Peru (1985)	2,249	8,050	0.99	0.026	0.042	0.030	287.56**
Peru (1991)	594	2,495	0.98	0.047	0.086	0.102	228.02**
South Africa (1993)	1,336	4,101	0.98	0.035	0.049	0.029	246.20**
South Africa (1995)	1,412	543	0.99	0.055	0.134	0.159	79.34**
South Africa (1998)	1,075	788	0.96	0.082	0.149	0.118	97.48**
Vietnam (1992)	4,800	20,936	0.99	0.034	0.049	0.045	762.36**
Vietnam (1998)	6,002	24,420	0.98	0.035	0.055	0.053	1092.93**

Note: \* indicates 95 percent and \*\* indicates 99 percent significantly different from Benford.

---

---

---

---

---

---

---

---

## Data editing

---

---

---

---

---

---

---

---

## Data editing and imputation

- The occurrence of nonresponse and errors makes it necessary to carry out an extensive process of checking the collected data, and, when necessary, correcting them.
- This checking and correction process is referred to as “**statistical data editing and imputation**”.



19

---

---

---

---

---

---

---

---

## To edit or not to edit?

a general principle – (de Waal et al. 2011: 13)

- The best editing technique is ... **no editing at all**, but instead ensuring that **correct data** are obtained during the **data collection phase**.
- To minimize errors, use **computer-assisted data collection**
- When an **invalid response** is given to a question during any of these data collection modes, this can be immediately reported by the computer that is used for data entry.
- The error can then be resolved by **asking** the respondent the relevant question(s) **again**.

20

---

---

---

---

---

---

---

---

## Computer-assisted data collection

- **Paper-based** personal interviewing (**PAPI**), coupled with computer-assisted **field-based** data entry (**CAFE**) was pioneered by the LSMS
- 1) Computer-Assisted **Personal** Interviewing (**CAPI**)
  - 2) Computer-Assisted **Telephone** Interviewing (**CATI**)
  - 3) Computer-Assisted **Self-Interviewing** (**CASI**)
  - 4) Computer-Assisted **Web** Interviewing (**CAWI**)

21

---

---

---

---

---

---

---

---

## Over-editing

a general principle

- It is not necessary to correct **all** data in **every detail**.
- **Small** errors in **individual records** are acceptable – they will not affect most of the statistics if interest.
- In order to obtain data of sufficiently high quality, it is usually enough to remove only the most **influential** errors.
- The boundary between **over-editing** and **creative editing** is blurred (de Waal et al. 2011: 3).

22

---

---

---

---

---

---

---

---

## Myanmar

Poverty and Living Conditions Survey 2015 (p.22)

- Before data entry: check of all the questionnaires for completeness of the data **coding mistakes**, logical links
- During data entry: batch editing - data checking including **skips**, **links** and data outliers
- After data entry: cleaning of the data - checks **out of range values**, violated skip patterns and logical links as well as **non-standard unit conversions**.



25

---

---

---

---

---

---

---

---

## Other errors

- **Missing data** are a special type of error and must be handled with great care
- Missing data are well-known troublemakers for both data producers (NSIs) and the analysts, and will therefore deserve a bit of our time today

26

---

---

---

---

---

---

---

---

## Missing data

27

---

---

---

---

---

---

---

## Afghanistan

Living conditions survey 2016-17, Percentage missing values for selected variables

Variable	Base population	Percent missing values
<b>Individual-level variables</b>		
Worked in business, organisation (12.2)	83,786	0.9
Person worked last month (12.6)	83,786	0.9
Days worked in past week (12.14)	34,772	2.1
Industry (12.16)	34,583	2.9
Place of birth (13.2)	155,680	0.7
Literacy (11.2)	121,829	0.9
Attended formal school (11.5)	121,829	0.8
Highest education grade completed (11.8)	52,422	2.1
Currently attending school (11.9)	41,178	1.7
Seeing disability (24.2)	155,680	0.3
Woman ever had a live birth (25.7)	17,633	3.6
Birth attendance (25.17)	17,633	4.6

28

---

---

---

---

---

---

---

## Missing data

- The occurrence of missing data is physiological
- Missing data imply loss of **precision** ('large' standard errors) and potential **bias** of the parameter estimates
- The loss of **precision** is a direct consequence of the smaller sample size implied by the presence of missing data
- The potential for **bias** is usually of far greater concern, and it all depends on the underlying **nonresponse mechanism**
- What do we mean by the underlying "nonresponse mechanism", exactly?

29

---

---

---

---

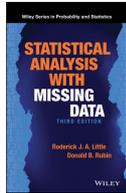
---

---

---

## A useful reference

Little and Rubin (2019)



### Definition

“Missing data are unobserved values that would be meaningful for analysis if observed; a missing value hides a meaningful value.”

### Mechanism

Why are data missing?

Different mechanisms lead to different strategies for treatment

---

---

---

---

---

---

---

---

## What causes missing data?

■ This is a difficult questions

■ While the reality is complex, we can grasp the essence of the discussion by focusing on two different mechanisms

1. “missing completely at random” (MCAR)
2. “missing not at random” (MNAR)

---

---

---

---

---

---

---

---

## MCAR

missing completely at random

- Technically, when missing data are MCAR, the **probability** that a value is missing **does not depend on the value** of the target variable or on the values of auxiliary variables.
- For example, data are MCAR when a respondent simply forgets to answer a question, or when a completed questionnaire is randomly lost during processing.
- Loosely, data are MCAR when data are missing **by accident**.
- Under MCAR, we have no reason to believe that missing data are different from observed data: the observed data can be regarded as a **random subset of the complete data**.

---

---

---

---

---

---

---

---

Unfortunately, MCAR rarely occurs in practice



33

---

---

---

---

---

---

---

---

### MNAR

missing not at random

- Technically, when missing data are MNAR, the **probability** that a value is missing **does depend on the value** of the target variable, and possibly also on the values of auxiliary variables.
- For example, data are MNAR when reported values of income are more likely to be missing for persons with a high income.
- The richer you are, the more reluctant you get to answer questions about your income.
- This situation is **the hardest to deal with analytically**, which is unfortunate because it may be the most likely.

34

---

---

---

---

---

---

---

---

### How to decide whether data are MCAR or MNAR?

- **Investigating the pattern of missingness** in the sample is paramount, before embarking on any action (data editing, dropping records, no action at all).
- Even **two-way tables** where the distribution of missing values is examined by region, urban-rural areas, and other dimensions are often insightful, despite their simplicity.
- When data are **MAR**, a useful rule of thumb is that **hot-deck** techniques are more suitable for elementary expenditure items, while **regression-based** imputation methods are a better option for large expenditure sub-aggregates.
- If there is evidence that data are **MNAR**, then the problem is serious, and ad-hoc procedures are required.

35

---

---

---

---

---

---

---

---

### How to deal with missing data?

- It depends.
- If data are **MCAR**, then observed data can be thought of as a random sample of the complete data, and statistical inference can be carried out based on "complete cases". Simply put, **missing data can be ignored**.
- If data are **MNAR**, the mechanism is referred to as non-ignorable missingness: observed data cannot be treated as if they were a random sample of the complete data. Standard estimation methods would produce biased estimates.

36

---

---

---

---

---

---

---

---

### To impute or not to impute?

- Methods to deal with missing data are countless
- Four categories:
  1. Procedures based on completely recorded units
  2. Weighting procedures
  3. Imputation
  4. Model-based methods

37

---

---

---

---

---

---

---

---

### Procedures based on completely recorded units

strategy I/IV

- **Strategy 1** consists in **discarding incomplete records** (that is, records with missing values) and analyzing only the units with complete data.
- For a data analyst, it is easy to implement and may be satisfactory with small amounts of missing data.
- What do you think?
- It can lead to serious **bias**, however, and it is **not efficient** (large standard errors).

38

---

---

---

---

---

---

---

---

## Weighting procedures

strategy II/IV

- Estimates from survey data are typically based on **Horvitz-Thompson (HT) estimators**.
- Take the **sample mean**, for example:

$$y_{HT} = \frac{1}{N} \sum_{i=1}^n w_i y_i$$

where  $n$  is the number of sampled units,  $N$  is the population size, and  $w_i = (\pi_i)^{-1}$  is the **weight** equal to the reciprocal of the probability of inclusion in the sample for unit  $i$ .

- Strategy 2** deals with missing data by **modifying the weights  $w_i$**  to adjust for different response rates. Similar to standard adjustment for unit nonresponse.

39

---

---

---

---

---

---

---

---

## Imputation

strategy III/IV

- Strategy 3** imputes the missing values, and analyzes the resultant completed data by standard methods.
- Popular imputation methods:
  - Hot deck imputation**  
each missing value is replaced with an observed value from a "similar" unit
  - Mean imputation**  
means from units with recorded values are substituted
  - Regression imputation**  
the missing variables for a unit are estimated by predicted values from the regression on the known variables for that unit

40

---

---

---

---

---

---

---

---

## Model-based methods

strategy IV/IV

- Strategy 4** - A broad class of procedures is generated by defining a **model** for the **complete data** and basing inferences on the likelihood distribution under that model
- Proper treatment of these methods would require an entire course...

41

---

---

---

---

---

---

---

---

Recap

- **Fatal errors**  
forewarned is forearmed
- **Systematic vs. random errors**  
randomness is preferable
- **How serious is the incidence of missing values?**  
Check and document
- **Is there any pattern in data missingness?**  
MCAR vs. MNAR
- **How to deal with missing values?**  
Should we ignore them? Or should we impute them?  
It depends on the mechanism (MCAR vs. MNAR)

C4D2 TRAINING 42

42

---

---

---

---

---

---

---

---

Data validation and diagnostics

C4D2 TRAINING 43

43

---

---

---

---

---

---

---

---

The question

- Even after editing and imputation, some errors typically remain in the datasets circulated and shared with analysts
- Analysts wonder if the datasets they receive qualify as a “high quality” dataset.
- **Data validation** provides an answer

C4D2 TRAINING 44

44

---

---

---

---

---

---

---

---

### Data validation: a working definition

- **Data validation** is a complex activity aimed at verifying that data intended to be used for analytical purposes are **cleaned** and **consistently organized** into datasets
- The **GIGO** principle  
"Garbage in, garbage out"

45

---

---

---

---

---

---

---

---

### Data validation: what is it?

1. **Range checks**  
simplest **edit** one can think of (details coming shortly)
2. **Internal consistency checks**  
combination of edits
3. **Outlier detection**  
investigation of extreme values
4. **Other data quality checks**

46

---

---

---

---

---

---

---

---

### Data validation: what is it?

1. **Range checks**  
simplest **edit** one can think of (details coming shortly)
2. **Internal consistency checks**  
combination of edits
3. **Outlier detection**  
investigation of extreme values (next lecture)
4. **Other data quality checks**

47

---

---

---

---

---

---

---

---

## Range checks

48

---

---

---

---

---

---

---

---

## Range checks

- **Range checks** aim at determining whether provided values are within allowable (legal) minima and maxima.
- Example: checking that the variable recording responses to question "Is ... female or male?" only takes values 1 or 2, which are the allowed (legal) response codes.
- Other examples?

49

---

---

---

---

---

---

---

---

## Examples

- 'age cannot be negative'
- 'profit must be smaller than or equal to revenue'
- **Your turn:** ..... must be ..... to .....
- **Your turn again:** ..... cannot be .....
- ...

50

---

---

---

---

---

---

---

---

## What if you come across an error?

Fellegi and Holt (1976: 17)

When a record fails some of the edits, we have, theoretically, five options:

1. Check the original questionnaires in the hope that the original questionnaire is correct and the edit failures are caused by coding error or error introduced during the conversion of data to machine-readable form;
2. Contact the original respondent to obtain the correct response (or verify that the reported response was correct in its original form);
3. Have clerical staff "correct" the questionnaire using certain rules which would remove the inconsistencies;
4. Use the computer to "correct" the questionnaire, also using certain rules which would remove the inconsistencies;
5. Drop all records which fail any of the edits or at least omit them from analyses using fields involved in failed edits.

51

---

---

---

---

---

---

---

---

## Internal consistency checks

52

---

---

---

---

---

---

---

---

## Internal consistency checks

- **Internal consistency checks** are intended to ensure that the information provided by the respondents contains no inconsistencies. Typically: 'in-record' (cross-variable) validation rules.

53

---

---

---

---

---

---

---

---

### Example

Namibia Household and Expenditure Survey  
2015/2016 report



Consistency between "Relationship to the head of the household" and "Marital status":  
if the relationship to the head of the household is "Spouse", then the individual must be married or in union; if there is a "Spouse/Partner" in the household, then the head of household must be married or in union.

54

---

---

---

---

---

---

---

---

### Other checks

55

---

---

---

---

---

---

---

---

### Other data quality checks

- A fourth type of checks can be designed and implemented with the aim of gaining insight into the overall data quality.
- Focus on two variables:
  - 1) the **age** reported by the households for each member
  - 2) the **expansion factors**

56

---

---

---

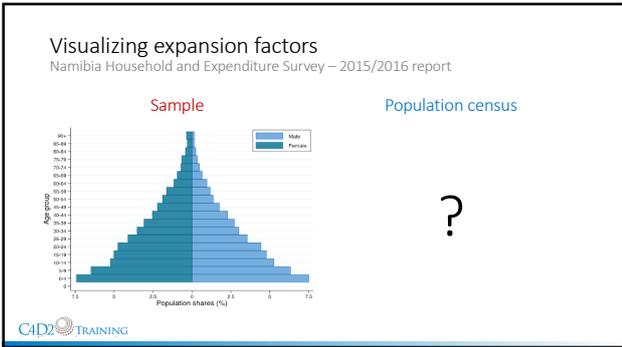
---

---

---

---

---



57

---

---

---

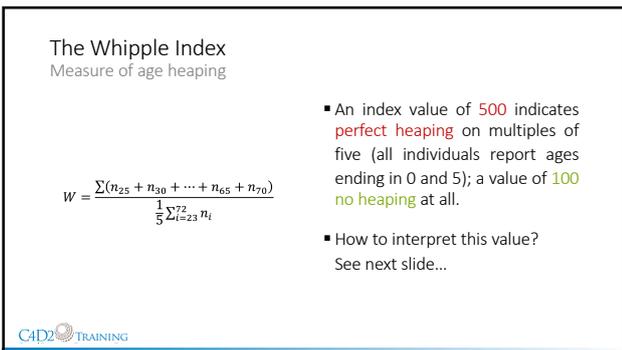
---

---

---

---

---



58

---

---

---

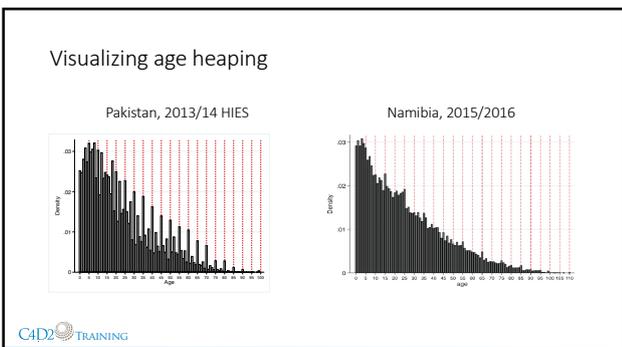
---

---

---

---

---



59

---

---

---

---

---

---

---

---

## United Nations Standards

Whipple's Index	Quality of data	Deviation from perfect
<105	Very accurate	<5%
105-110	Relatively accurate	5-9.99%
110-125	OK	10-24.9%
125-175	Bad	25-74.99%
>175	Very bad	>=75%

Two questions:

- 1) What do we know about **other countries**?
- 2) How confident are we about the **interpretation** of the Whipple index as a summary measure of data quality?

60

---

---

---

---

---

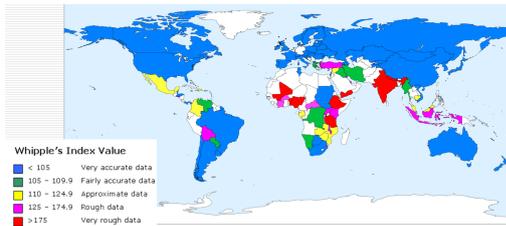
---

---

---

## Short answers

What do we know about other countries?



61

---

---

---

---

---

---

---

---

## How confident are we about this interpretation of the Whipple index?

- What if, rather than data quality, Whipple's index captures **literacy, numeracy, incentives...** ultimately, **living standards**?
- Areas where school enrolment is low, illiteracy is high, income is low, are typically those where age heaping is more pronounced
- Incentives** to remember one's exact age may be low: people in poor, isolated, rural areas may have little use for it in their everyday lives
- That doesn't always mean they cannot respond accurately to other questions

62

---

---

---

---

---

---

---

---



## Lessons learned

- Errors are ubiquitous: data validation techniques are meant to localize and fix errors.
- A special type of error is missing data.
- The key issue is whether the reasons for missingness are related to the variable of interest. When data are **MCAR**, the impact of missing data is relatively benign. When data are **MNAR**, then ignoring missing data would lead to biased estimates.
- Imputation of missing values – best method depends on the nonresponse mechanism.
- Data validation and diagnostic should be routine.
- Proper **documentation** of the validation process is an integral part of the metadata to be published.

63

---

---

---

---

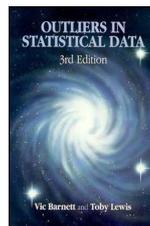
---

---

---

---

## Useful references



64

---

---

---

---

---

---

---

---

## References

### Required readings

De Waal, T., Pannekoek, J., and Scholtus, S. (2011). Handbook of Statistical Data Editing and Imputation. New York: John Wiley and Sons (Chapter 1)

### Suggested readings

Barnett, V., and Lewis T. (1994). Outliers in Statistical Data. 3rd edition. J. Wiley & Sons 1994, XVII, 582 pp.

Dang, H. A., Jolliffe, D., & Carletto, C. (2018). Data Gaps, Data Incomparability, and Data Imputation. Ecineq WP, 456.

Fellegi, I. P., & Holt, D. (1976). A systematic approach to automatic edit and imputation. Journal of the American Statistical association, 71(353), 17-35.

Harrison, D. E., & Krauss, S. I. (2002). Interviewer cheating: Implications for research on entrepreneurship in Africa. Journal of Developmental Entrepreneurship, 7(3), 319.

Judge, G., & Schechter, L. (2009). Detecting problems in survey data using Benford's Law. Journal of Human Resources, 44(1), 1-24.

Little, R. J., & Rubin, D. B. (2019). Statistical analysis with missing data (Vol. 793). Wiley.

65

---

---

---

---

---

---

---

---

Thank you for your attention

C4D2 TRAINING 66

66

---

---

---

---

---

---

---

Homework

C4D2 TRAINING 67

67

---

---

---

---

---

---

---

Exercise 1 – Engaging with the literature



- Finn and Ranchhod (2017) study the implications of data fabrication in practice, using the case of South Africa.
- Read the paper and summarize its main methods and findings.

C4D2 TRAINING 68

68

---

---

---

---

---

---

---

## Exercise 2 – Methods for detecting ‘fakes’

- Summarize the main findings of Bredl et al (2012), and compare their method with Finn and Ranchhod (2017)

Survey Methodology, June 2012  
DOI: 10.1002/sur.1005  
Statistica Canada, Catalogue No. 12-001-X

1

### A statistical approach to detect interviewer falsification of survey data

Sebastian Bredl, Peter Winker and Kerstin Klotzsch\*

#### Abstract

Survey data are potentially affected by interviewer falsifications with data falsification being the most frequent form. There is a small number of falsification detection methods available, which are based on the analysis of the data. This paper discusses how these methods can be used to detect interviewer falsification. The paper also discusses how these methods can be used to detect interviewer falsification. The paper also discusses how these methods can be used to detect interviewer falsification.

Key Words: Data falsification; interviewer falsification; cluster analysis.

69

---

---

---

---

---

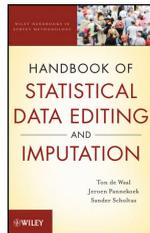
---

---

---

## Exercise 3 – Computer-assisted data collection modes

- Summarize the discussion of the pros and cons of the four modes of data collection in De Waal et al. 2011, p. 14.



70

---

---

---

---

---

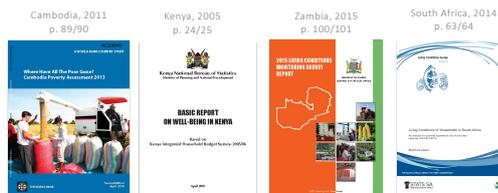
---

---

---

## Exercise 4 – Data imputation – I/II

- Review the imputation procedures described in the following reports:



71

---

---

---

---

---

---

---

---

#### Exercise 4 – Data imputation – II/II

- How did these countries deal with missing data and outliers?
- Based on the lessons learned in this lecture, what is your opinion about each method used?

- Links to the reports:

Cambodia 2011: <http://documents.worldbank.org/curated/en/824314680174055777/where-have-all-the-poor-gone-cambodia-slow-the-paces-slowly-2013>  
Kenya 2005: <http://statistics.tbsa.or.ke/nada/index.php/catalog/8>  
Zambia 2015: [https://www.zamstats.gov.zm/nhocaldownload/ivime\\_Conditions/2015%20ivime%20Conditions%20Monitoria%20Curv%20Report.pdf](https://www.zamstats.gov.zm/nhocaldownload/ivime_Conditions/2015%20ivime%20Conditions%20Monitoria%20Curv%20Report.pdf)  
South Africa 2014: <http://www.statssa.gov.za/publications/P0310/P03102014.pdf>

---

---

---

---

---

---

---

---