Data validation and diagnostics

LECTURE 11

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.

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)

Types of data errors

Types of errors

Errors can be classified in several ways:

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

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.

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

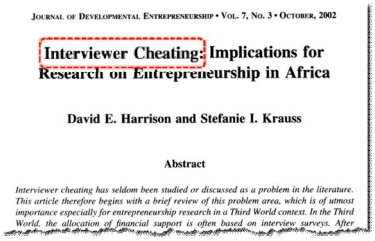
Fatal errors

- 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?

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.

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

C4D2 TRAINING

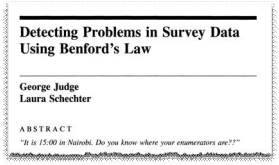
Prevention

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

C4D2 TRAINING

Ruling out horror stories



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??"

Judge and Schechter (2009)

- "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.

C4D2 TRAINING

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).

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:

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

- 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.

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.
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

C4D2 Training

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.

C4D2 TRAINING

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

	Surveys	Observations	r	m	d^*	a^*	χ^2	V_K^*
Bangladesh (1996)	4,522	7,357	0.92	0.065	0.085	0.039	440.92**	8.27**
Ghana (1991/92)	294	799	0.99	0.006	0.026	0.010	8.45	0.85
Mexico (Oct 90)	24,067	10,028	0.84	0.154	0.169	0.059	3959.21**	15.45**
Mexico (May 91)	22,328	8,659	0.88	0.134	0.155	0.053	3145.66**	13.14**
Pakistan (Jul 86)	927	874	0.97	0.060	0.058	0.036	31.62**	1.84**
Pakistan (Oct 86)	909	181	0.86	0.100	0.128	0.016	23.37*	1.40*
Pakistan (Jan 87)	882	563	0.89	0.047	0.062	0.046	20.71*	1.13
Pakistan (Mar 87)	854	46	0.94	0.051	0.090	0.028	4.79	0.73
Pakistan (Apr 87)	845	191	0.96	0.041	0.078	0.059	15.90*	1.13
Pakistan (July 87)	831	658	0.97	0.033	0.061	0.022	43.75**	0.95
Pakistan (Jan 88)	813	728	0.96	0.046	0.069	0.001	41.45**	1.41*
Pakistan (Mar 88)	809	69	0.92	0.048	0.096	0.050	8.11	1.16
Pakistan (Aug 88)	804	379	0.99	0.022	0.030	0.027	13.44	0.77
Pakistan (Jan 89)	802	708	0.99	0.014	0.032	0.054	26.23**	1.27
Pakistan (Mar 89)	766	200	0.78	0.134	0.141	0.057	29.67**	2.11**
Pakistan (Aug 89)	759	691	0.98	0.025	0.045	0.020	22.18*	1.06
Pakistan (Oct 91)	726	1,095	0.99	0.022	0.033	0.017	18.72*	1.16
Paraguay (1989)	298	1,412	0.99	0.026	0.036	0.025	30.36**	1.21
Paraguay (2000)	223	1,632	0.97	0.050	0.065	0.051	101.34**	2.69**
Peru (1985)	2,349	8,050	0.99	0.026	0.042	0.030	287.76**	4.23**
Peru (1994)	591	5,105	0.98	0.017	0.064	0.102	228.02**	5.17**
Peru (1998)	1,336	4,101	0.98	0.036	0.065	0.059	246.39**	4.63**
South Africa (1993)	1,412	543	0.99	0.087	0.134	0.159	79.34**	3.89**
South Africa (2001)	1,075	788	0.96	0.083	0.108	0.118	97.48**	3.68**
Vietnam (1992)	4,800	20,936	0.99	0.055	0.049	0.015	762.26**	7.80**
Vietnam (1998)	6,002	24,420	0.98	0.035	0.055	0.053	1092.93**	9.64**

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

C4D2 TRAINING

Data editing

C4D2 TRAINING

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".

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.

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)

Over-editing
a general principle

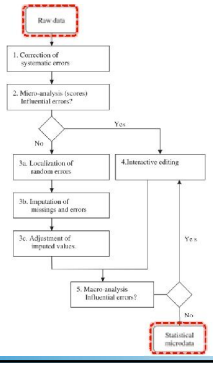
- It is not necessary to correct all data in every detail.
- Small errors in individual records are acceptable – they will not affect most of the statistics if interest.
- In order to obtain data of sufficiently high quality, it is usually enough to remove only the most influential errors.
- The boundary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

Modern editing methods

- Interactive editing
  Specialist-assisted check of specified edits during or after data entry
- Selective editing
  Split the data into a 'critical stream' (records likely to contain influential errors) and a 'noncritical stream'
- Macro-editing
  - aggregation method: verify if figures to be published seem plausible
  - distribution method: available data are used to construct the distribution of the variables and uncommon values are candidates for further inspection or editing
- Automatic editing
  records are edited by a computer without human intervention

1
Identify and fix errors that are evident and easy to treat with sufficient reliability

2
Select records for interactive editing that contain influential errors that cannot be treated automatically with sufficient reliability.
Use scores.



3
Apply all relevant automatic editing to the (many) records that are not selected for interactive editing in step 2.

4
Apply interactive editing to the minority of the records with influential errors.

5
Macro editing on records with influential errors.

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.

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

Missing data

Afghanistan

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

Variable	Base population	Percent missing values
Individual-level variables		
Worked in business, organisation (12.2)	83,788	0.9
Person worked last month (12.6)	83,788	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	0.6
Highest education grade completed (11.6)	52,422	2.1
Currently attending school (11.9)	41,173	1.7
Seeing disability (24.2)	155,680	0.3
Woman ever had a live birth (25.7)	17,163	3.6
Birth attendance (25.17)	17,163	4.6

C4D2 TRAINING

28

Missing data

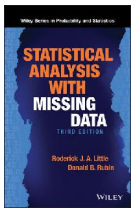
- The occurrence of missing data is physiological
- Missing data imply loss of precision ('large' standard errors) and potential bias of the parameter estimates
- The loss of precision is a direct consequence of the smaller sample size implied by the presence of missing data
- The potential for bias 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?

C4D2 TRAINING

29

A most 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

C4D2 TRAINING

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

- Data are MCAR when, for instance, a respondent forgets to answer a question or when a random part of the data is lost while processing it.
- 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.
- 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.

*MCAR

formal definition from De Waal et al. (2011)

More formally, a nonresponse mechanism is called MCAR if

$$P(r_j \mid y_j, \mathbf{x}, \xi) = P(r_j \mid \xi). \quad (1.1)$$

In this notation, r_j is the response indicator of target variable y_j, where $r_{ij} = 1$ means that record i contains a response for variable y_j, and $r_{ij} = 0$ that the value of variable y_j is missing for record i, $\mathbf{x}$ is a vector of always observed auxiliary variables, and ξ is a parameter of the nonresponse mechanism.

*MAR

missing at random, definition from De Waal et al. (2011)

In more formal terms, a nonresponse mechanism is called MAR if

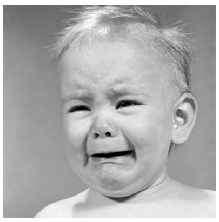
$$P(r_j \mid y_j, \mathbf{x}, \xi) = P(r_j \mid \mathbf{x}, \xi), \tag{1.2}$$

using the same notation as in (1.1).

MAR is a more complicated situation than MCAR. In the case of MAR, one needs to find appropriate groups of population units to reduce MAR to MCAR for these groups. Once these groups of population units have been found, it is simple to correct for missing data because within these groups all units may be assumed to have the same probability to respond.

In practice, one usually assumes the nonresponse mechanism to be MAR and tries to construct appropriate groups of population units. These groups are then used to correct for missing data.

Unfortunately, MCAR rarely occurs in practice



MNAR

missing not at random

- Data are MNAR when, for instance, reported values of income are more likely to be missing for persons with a high income.
- 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.
- This situation is the hardest to deal with analytically, which is unfortunate because it may be the most likely.

*MNAR

formal definition from De Waal et al. (2011)

In more formal terms, a nonresponse mechanism is called NMAR if

$$P(r_j \mid y_j, \mathbf{x}, \xi)$$

cannot be simplified, that is, if both (1.1) and (1.2) do not hold.

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 produced biased estimates.

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

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).

Weighting procedures

strategy II/IV

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

$$\bar{y}_{HT} = \sum_{i=1}^{n} \pi_i^{-1} y_i / N$$

- where n is the number of sampled units, N is the population size, and π_i is the known probability of inclusion in the sample for unit i.
- Strategy 2 deals with missing data by modifying the weights to adjust for different response rates. Similar to standard adjustment for unit nonresponse.

*More on weighting procedures

De Waal et al. (2011)

$$\bar{y}_{\mathrm{HK}} = \frac{\sum_{i=1}^{n} \pi_i^{-1} y_i}{\sum_{i=1}^{n} \pi_i^{-1}}, \tag{1.7}$$

where the sums are over the n sampled units, N is the population size, and π_i is the known probability of inclusion in the sample for unit i. Weighting procedures for nonresponse modify the weights in an attempt to adjust for nonresponse as if it were part of the sample design. The resultant estimator (1.7) is replaced by

$$\frac{\sum_{i=1}^{n} (\pi_i \hat{p}_i)^{-1} y_i}{\sum_{i=1}^{n} (\pi_i \hat{p}_i)^{-1}},$$

where the sums are now over sampled units that respond, and $\hat{p}_i$ is an estimate of the probability of response for unit i, often the proportion of responding units in the subclass of the sample in which unit i falls. Weighting methods are described further in Chapter 3.

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

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

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 the 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)

Data validation and diagnostics

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 help to answer

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"

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

C4D2 Training

Range checks

C4D2 Training

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?

C4D2 Training

Examples

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

Possible outcomes

- Error
  the data are rejected.
- Warning
  the data can be accepted, with some corrections or explanations from the data provider;
- Information
  the data are accepted. No error.

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.

Internal consistency checks

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.

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.

Other checks

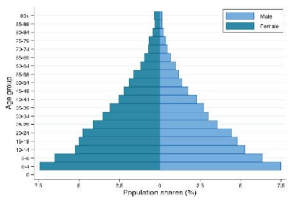
Other data quality checks

- 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

Visualizing expansion factors

Namibia Household and Expenditure Survey – 2015/2016 report

Sample



Population census

?

The Whipple Index

Measure of age heaping

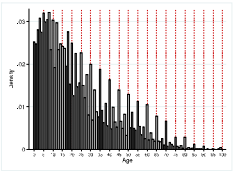
$$W = \frac{\sum(n_{25} + n_{30} + \cdots + n_{65} + n_{70})}{\frac{1}{5}\sum_{i=23}^{72} n_i}$$

- An index value of 500 indicates perfect heaping on multiples of five (all individuals report ages ending in 0 and 5); a value of 100 no heaping at all.
- What do we get for the 2013/14 HIES data?
- How to interpret this value? See next slide…

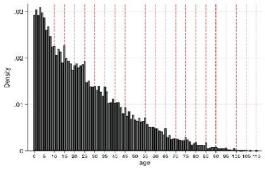
C4D2 Training

Visualizing age heaping

Pakistan, 2013/14 HIES



Namibia, 2015/2016



C4D2 Training

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%

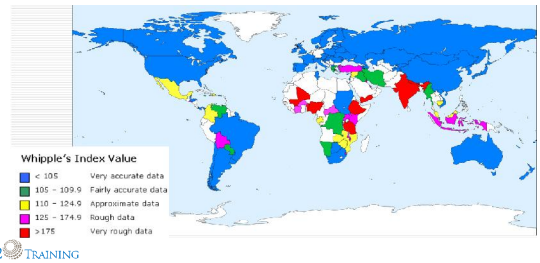
- Two questions:

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

C4D2 Training

Short Answers

What do we know about other countries?



C4D2 TRAINING

How confident we are about the interpretation of the Whipple index as a summary measure of data quality?

- What if, rather than data quality, Whipple's index captures literacy, numeracy… ultimately, living standards?

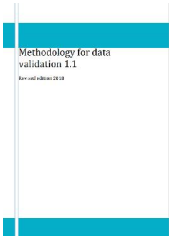
C4D2 TRAINING

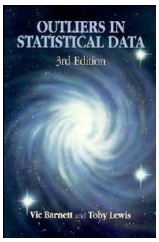
Lessons learned

- 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 outcome 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
- A proper documentation of the validation process is an integral part of the metadata to be published

C4D2 TRAINING 66

Useful references





C4D2 TRAINING

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.

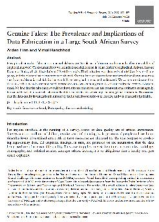
C4D2 TRAINING

Thank you for your attention

C4D2 TRAINING

Homework

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.

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 1
Vol. 38, No. 1, pp. 1-10
Statistics Canada, Catalogue No. 12-001-X

A statistical approach to detect interviewer falsification of survey data

Sebastian Bredl, Peter Winker and Kerstin Kötschau [1]

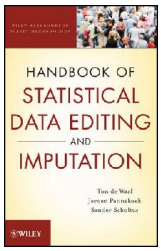
Abstract

Survey data are potentially affected by interviewer falsifications with data fabrication being the most blatant form. Even a small number of fabricated interviews might seriously impair the results of further empirical analysis. Besides reinterviews, some statistical approaches have been proposed for identifying this type of fraudulent behaviour. With the help of a small dataset, this paper demonstrates how cluster analysis, which is not commonly employed in this context, might be used to identify interviewers who falsify their work assignments. Several indicators are combined to classify 'at risk' interviewers based solely on the data collected. This multivariate classification seems superior to the application of a single indicator such as Benford's law.

Key Words: Data fabrication; Falsifier; Benford's law; Cluster analysis.

Exercise 3 – Computer-assisted data collection modes

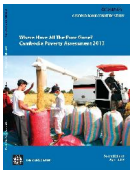
- Summarize the discussion of the pros and cons of the four modes of data collection in De Waal et al. 2011, p. 14.



Exercise 4 – Data imputation – I/II

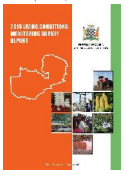
- Review the imputation procedures described in the following reports:

Cambodia, 2011
p. 89/90



Kenya, 2005
p. 24/25

Zambia, 2015
p. 100/101



South Africa, 2014
p. 63/64

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/824341468017405577/Where-have-all-the-poor-gone-Cambodia-poverty-assessment-2013
  - Kenya 2005: http://statistics.knbs.or.ke/nada/index.php/catalog/8
  - Zambia 2015: https://www.zamstats.gov.zm/phocadownload/Living_Conditions/2015%20Living%20Conditions%20Monitoring%20Survey%20Report.pdf
  - South Africa 2014: http://www.statssa.gov.za/publications/P0310/P03102014.pdf