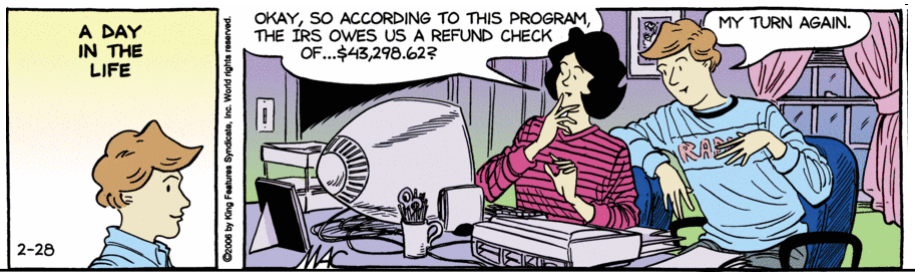
Data Cleaning & Checking: Minimising Garbage

Prof. Gavin T L Brown (gt.brown@auckland.ac.nz)
The University of Auckland
Lecture notes on research methods.

THE UNIVERSITY OF AUCKLAND
FACULTY OF EDUCATION
Te Kura Akoranga o Tāmaki Makaurau
Incorporating the Auckland College of Education

What's the point?

- Quality of inferences depends on KNOWING that the data being analysed are a true and accurate record of reality and that they represent what you think they are supposed to
- NOT wanted➔GIGO

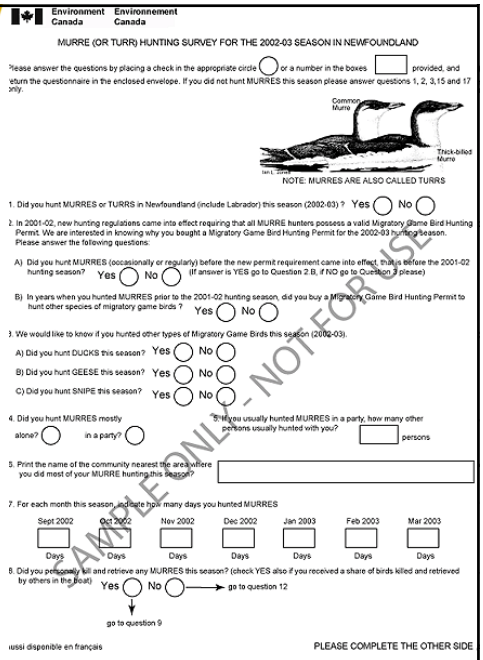


Things that go bump in the night

- Wrong values
- Response sets
- Jokesters
- Impossible values
- Missing values
- Extreme values

Wrong Values

- Check Sample of Data Entry cases against SOURCE documents
  - 10% systematic sample to start
  - If all values correct, then proceed
  - If values wrong, check ALL
  - Be sure that the digital file represents accurately the source



Environment Canada / Environnement Canada

MURRE (OR TURR) HUNTING SURVEY FOR THE 2002-03 SEASON IN NEWFOUNDLAND

Please answer the questions by placing a check in the appropriate circle ◯ or a number in the boxes ☐ provided, and return the questionnaire in the enclosed envelope. If you did not hunt MURRES this season please answer questions 1, 2, 3,15 and 17 only.

Common Murre — Thick-billed Murre

NOTE: MURRES ARE ALSO CALLED TURRS

1. Did you hunt MURRES or TURRS in Newfoundland (include Labrador) this season (2002-03) ? Yes ◯ No ◯

2. In 2001-02, new hunting regulations came into effect requiring that all MURRE hunters possess a valid Migratory Game Bird Hunting Permit. We are interested in knowing why you bought a Migratory Game Bird Hunting Permit for the 2002-03 hunting season. Please answer the following questions:

A) Did you hunt MURRES (occasionally or regularly) before the new permit requirement came into effect, that is before the 2001-02 hunting season? Yes ◯ No ◯ (If answer is YES go to Question 2.B, if NO go to Question 3 please)

B) In years when you hunted MURRES prior to the 2001-02 hunting season, did you buy a Migratory Game Bird Hunting Permit to hunt other species of migratory game birds ? Yes ◯ No ◯

3. We would like to know if you hunted other types of Migratory Game Birds this season (2002-03).

A) Did you hunt DUCKS this season? Yes ◯ No ◯

B) Did you hunt GEESE this season? Yes ◯ No ◯

C) Did you hunt SNIPE this season? Yes ◯ No ◯

4. Did you hunt MURRES mostly alone? ◯ in a party? ◯

5. If you usually hunted MURRES in a party, how many other persons usually hunted with you? ☐ persons

5. Print the name of the community nearest the area where you did most of your MURRE hunting this season?

7. For each month this season, indicate how many days you hunted MURRES

Sept 2002	Oct 2002	Nov 2002	Dec 2002	Jan 2003	Feb 2003	Mar 2003
☐ Days	☐ Days	☐ Days	☐ Days	☐ Days	☐ Days	☐ Days

8. Did you personally kill and retrieve any MURRES this season? (check YES also if you received a share of birds killed and retrieved by others in the boat) Yes ◯ No ◯ → go to question 12

go to question 9

ussi disponible en français

PLEASE COMPLETE THE OTHER SIDE

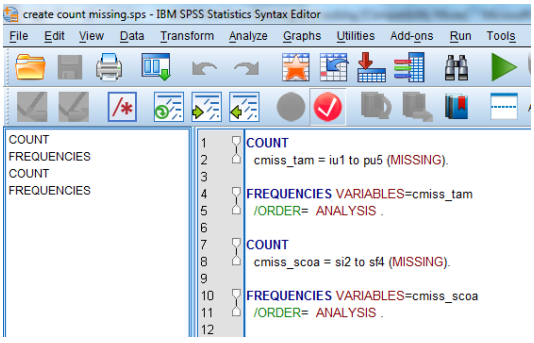
SAMPLE ONLY - NOT FOR USE

Response set

- Biased way of responding that invalidates data
  - If unwilling, then may be careless/hasty
  - If unwilling, then may deliberately mislead
  - If trouble deciding, then may guess or choose socially desirable response
- Look for
  - All answers the same—clearly invalid
  - A physical pattern of responses on the page
  - Compare logically opposite items; if same answer then maybe responses not valid
  - Jokesters: Fan, X., Miller, B. C., Park, K.-E., Winward, B. W., Christensen, M., Grotevant, H. D., & Tai, R. H. (2006). An Exploratory Study about Inaccuracy and Invalidity in Adolescent Self-Report Surveys. *Field Methods 18(3), 223-244. doi:10.1177/152822X06289161*

Count Missing responses

- Select range of items to check—inventory specific.
- Reject cases with >10% missing
- Mark each case as to whether it is kept or not

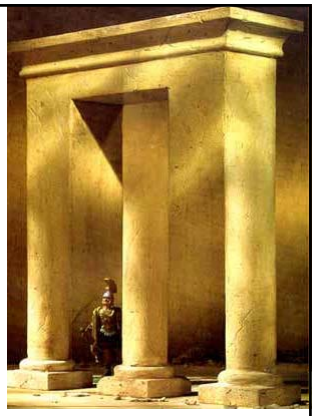


cmiss_scoa

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	198	89.2	89.2	89.2
	1.00	17	7.7	7.7	96.8
	2.00	6	2.7	2.7	99.5
	3.00	1	.5	.5	100.0
	Total	222	100.0	100.0	

Impossible Values

- Check Minimum & Maximum are valid
  - Cannot be higher or lower than allowed
- Check all responses are valid codes
  - 0 is not a code, it is a value
  - Missing response should be obvious arbitrary code (e.g., -9)
- Check logic of inter-linked responses
  - e.g., If Year 8, age<14; if school=intermediate, year=7 or 8 only; if sex=F, single sex school≠Male; etc.
  - Maximum count is 100%



Imagine a strange missing pattern

Case	1	2	3	4	5	6	7	8	9
A	.	2	3	4	5	1	2	3	4
B	1	.	3	4	5	1	2	3	4
C	3	4	.	4	5	1	2	3	4
D	3	4	4	.	1	2	3	4	5
E	3	4	4	1	.	1	2	3	4
F	4	5	4	1	5	.	2	4	5
G	3	4	5	1	5	4	.	1	3
H	4	5	4	2	5	4	3	.	2
I	1	2	5	1	1	4	3	2	.

Any analysis that requires all people to answer all items will fail even though each person is missing only 1 answer

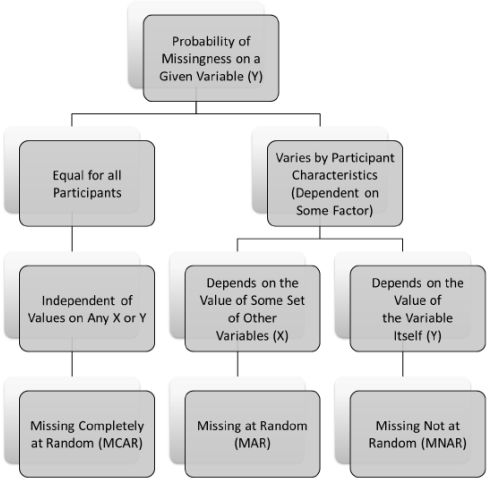
Missing Values

- Too much missing
  - >10%➔delete case/variable
- A little missing
  - <10% within tolerance
  - Goal: prevent listwise dropping of otherwise valid cases

Types of missing data

Source:

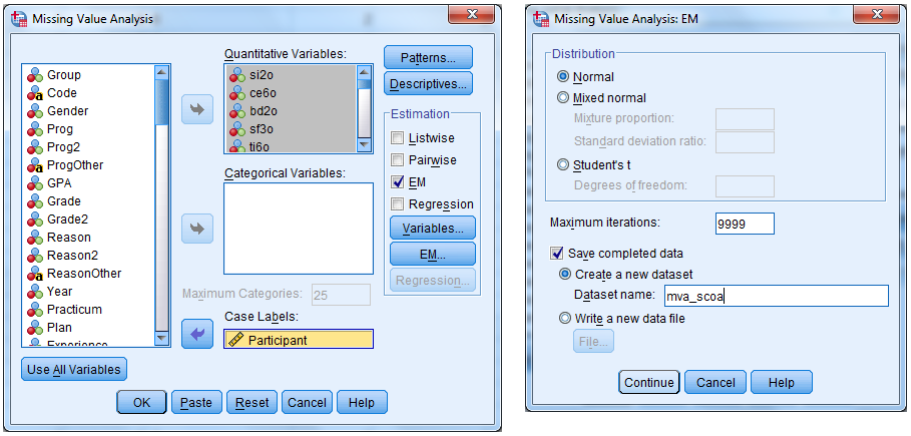
Teresa A. Myers (2011). Goodbye, listwise deletion: Presenting Hot Deck Imputation as an easy and effective tool for handling missing data. *Communication Methods & Measures*, *5*(4), 297-310



Expectation Maximisation

- Impute missing with EM procedure
  - EM uses MLE to check that M, SD, correlations, covariances not disturbed by imputation
  - Assumption is that the sample input values are the best estimate of the population values
    - Requires sampling to be high quality
  - Iteratively imputes values and checks which values disturb resulting matrices least
  - **PS** check descriptives and MCAR test post-imputation to be sure EM variables are ok to use

EM Missing Value Analysis—Setup



MVA: EM

- Check the % missing per variable.
- IF <10% proceed, otherwise delete variable.

Univariate Statistics

	N	Mean	Std. Deviation	Missing Count	Missing Percent	No. of Extremes[a] Low	No. of Extremes[a] High
si2o	222	4.15	1.353	0	.0	0	0
ce6o	222	3.77	1.289	0	.0	0	0
bd2o	222	2.55	1.242	0	.0	0	19
sf3o	222	3.15	1.231	0	.0	0	0
ti6o	219	3.77	1.221	3	1.4	0	0
pe2o	221	3.21	1.189	1	.5	0	0
ig3o	221	2.15	1.189	1	.5	0	1
ti5o	220	4.01	1.137	2	.9	0	0
ti4o	222	3.82	1.140	0	.0	0	0
si3o	221	4.08	1.165	1	.5	21	0
sq2o	222	3.52	1.250	0	.0	13	12
ce3o	221	3.78	1.144	1	.5	0	0
bd5o	222	3.26	1.320	0	.0	0	0
si1o	222	4.09	1.064	0	.0	0	0
si4o	222	3.95	1.096	0	.0	0	0
sf2o	222	3.50	1.351	0	.0	24	16
ce4o	222	3.44	1.178	0	.0	12	5
bd4o	222	2.85	1.169	0	.0	0	25
si5o	221	3.75	1.039	1	.5	6	10
sf1o	220	3.64	1.339	2	.9	0	0
ce1o	219	3.32	1.107	3	1.4	8	6
bd3o	221	3.43	1.161	1	.5	4	12
ti1o	218	3.72	1.052	4	1.8	2	10
sq1o	220	3.28	1.116	2	.9	11	5
ce5o	218	3.35	1.131	4	1.8	12	4
bd1o	222	2.57	1.158	0	.0	0	12
ti2o	221	3.55	1.153	1	.5	5	16
ce2o	222	3.66	1.145	0	.0	6	10
ig2o	221	2.23	1.256	1	.5	0	5
ti3o	220	3.96	1.029	2	.9	0	0
pe1o	222	3.11	1.259	0	.0	0	0
ig1o	220	2.63	1.196	2	.9	0	16
sf4o	222	3.85	1.044	0	.0	3	11

a. Number of cases outside the range (Q1 - 1.5*IQR, Q3 + 1.5*IQR).

Checking MVA effects

- How large a difference did imputation make to M and SD?
  - Usually 2nd & 3rd decimal point

Summary of Estimated Means

	si2o	ce6o	bd2o	sf3o	ti6o	pe2o	ig3o	ti5o	ti4o	si3o	sq2o	ce3o	bd5o	si1o	si4o	sf2o	ce4o	bd4o	si5o
All Values	4.15	3.77	2.55	3.15	3.77	3.21	2.15	4.01	3.82	4.08	3.52	3.78	3.26	4.09	3.95	3.50	3.44	2.85	3.75
EM	4.15	3.77	2.55	3.15	3.78	3.21	2.15	4.01	3.82	4.07	3.52	3.77	3.26	4.09	3.95	3.50	3.44	2.85	3.75

	si2o	ce6o	bd2o	sf3o	ti6o	pe2o	ig3o	ti5o	ti4o	si3o	sq2o	ce3o	bd5o
All Values	1.353	1.289	1.242	1.231	1.221	1.189	1.189	1.137	1.140	1.165	1.250	1.144	1.320
EM	1.353	1.289	1.242	1.231	1.220	1.189	1.187	1.134	1.140	1.180	1.250	1.144	1.320

Validity of Imputation

- Distribution of missing should be random
- EM provides Little's X2 test of Missing Completely at Random (MCAR)
  - Missing value not dependent on any other variable
- When in doubt divide χ^2/df and look up the stat sig of that value. See http://www.fourmilab.ch/rpkp/experiments/analysis/chiCalc.html

si2o	ce6o	bd2o	sf3o	ti6o	pe2o	ig3o	ti5o	ti4o	si3o	sq2o	ce3o	bd5o	si1o
4.15	3.77	2.55	3.15	3.78	3.21	2.15	4.01	3.82	4.07	3.52	3.77	3.26	4.09

a. Little's MCAR test: Chi-Square = 634.015, DF = 507, Sig. = .000

χ^2/df=1.25; p=.26

Check the imputation for possible invalid imputations

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
si2	220	.97	6.00	4.1296	1.34736	-.783	.164	.015	.327
ce6	220	1.00	6.00	3.7607	1.28966	-.381	.164	-.318	.327
bd2	220	1.00	6.00	2.5409	1.24716	.911	.164	.432	.327
sf3	220	1.00	6.00	3.1364	1.22364	-.067	.164	-.609	.327
ti6	220	1.00	6.00	3.7667	1.21688	-.343	.164	-.456	.327
pe2	220	1.00	6.00	3.2018	1.19054	-.117	.164	-.571	.327
ig3	220	1.00	6.00	2.1648	1.18609	1.051	.164	.432	.327

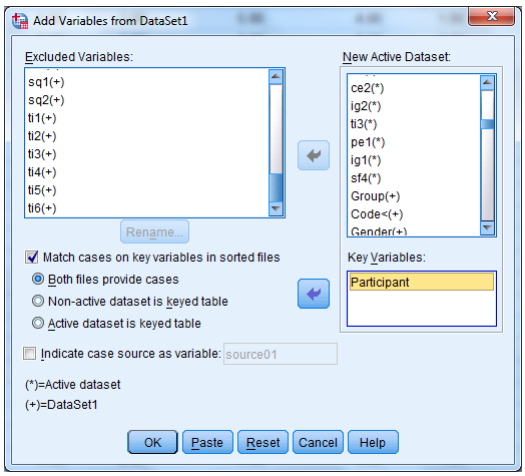
Find the offending case (sort ascending or descending)
Correct it to valid min or max value
Use these values

Import imputed values back into master data file

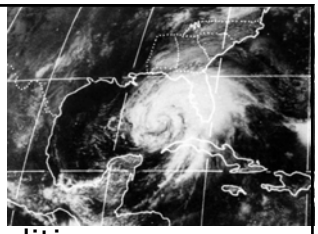
- Use data merge procedure but
  - Rename variables so that they have slightly different file names. For example
    - add an o for original to the original var
    - Add an m for missing to the new var
  - Put data in ascending order for the key variable
    - Unique identifier that you used

Merge variables

- Run Merge <add variables>
- Match files using key variable



Extreme Values



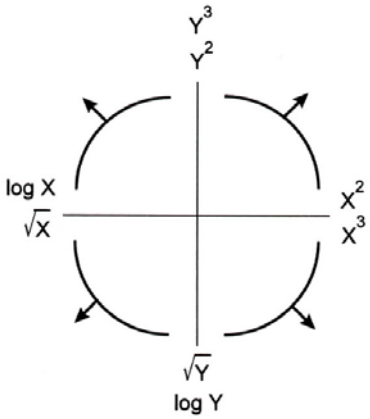
- Do not represent well normal conditions
  - Mean is very sensitive to extreme values
  - Need to detect and resolve (adjust or delete)
- Outlier detection
  - Check kurtosis & skewness
    - (+/-3.0 no problem)+in some cases as high as 7.00 is ok
  - Check boxplot displays for people with extreme values per variable

Dealing with non-normality

- Remove
- Robustify (adjust using a trimming technique)
  - Use Median or median absolute difference to substitute for Mean and SD if outliers present
  - Huber's method or winsorise:
    - 90% Winsorised mean sets the bottom 5% to the 5th percentile value, the top 5% to the 95th percentile value, and then evaluates the variable for normality—repeat until normal.
  - http://www.rsc.org/images/brief6_tcm18-25948.pdf

Dealing with non-normality

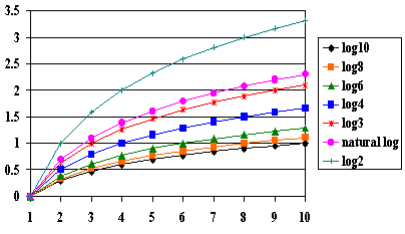
- Transform (multiply by a constant to make normal or linear)
  - Bulging rule—depending on shape of distribution try these transformations to make variable linear
    - Mosteller, Frederick, & Tukey, John W. (1977). *Data Analysis and Regression: A Second Course in Statistics. Reading, MA: Addison-Wesley.*



Dealing with Non-normality

- Square root transformation.
  - Add constant so min=2.00
- Log transformation(s).
  - Add constant so min=1.00
- Inverse transformation.
  - After *-1, add constant so min = 1.00
- **Beware**: transformations improve normality, but curvilinear transformations affect interpretation of results

Figure 1.
The Effect of log base on the efficacy of transformations.



Osborne, J. W. (2002). Notes on the use of data transformations. *Practical Assessment, Research & Evaluation, 8(6), PAREonline.net/getvn.asp?v=8&n=6.*

Box-Cox transformation for non-normality

1. assess variable to find the optimal power transformation (λopt).
   - Use online software produced by Wessa (2013)
2. add/subtract constant (c) to make variable min = 1.00
3. transform each value: $(x +/- c)^{\lambda opt}$
   - Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, B26*, 211-234.
   - Osborne, J. (2010). Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research & Evaluation, 15*(12), http://pareonline.net/pdf/v15n12.pdf.
   - Wessa, P. (2013). *Box-Cox Normality Plot—Free Statistics Software. Office for Research Development and Education, version 1.1.23-r7*. Retrieved from http://www.wessa.net/rwasp_boxcoxnorm.wasp

When in doubt

- Test the transformation by conducting *Sensitivity analysis*
  - Run the analysis using the original and transformed values
  - Evaluate the results for the substantive impact
- Example from Osborne 2010
  - correlation between number of faculty (many small universities, few large ones) and associate professor salary (before transformation) $r_{(1161)} = 0.49$, $p < .0001$. (% variance accounted for =0.24)
  - After optimal transformation, $r_{(1161)} = 0.66$, $p < .0001$. % variance accounted for = 0.44 (an 81.5% increase)
  - *Which is correct? Make the argument for the better result*

Support material

- http://www.tulane.edu/~panda2/Analysis2/datclean/dataclean.htm
- http://www.amstat.org/publications/jse/v13n3/datasets.holcomb.html#Mason
- Robson, C. (2002). *Real World Research* (2nd ed.) (pp. 391-398). Oxford: Blackwell.
- McClelland, G. H. (2000). Nasty data: Unruly, ill-mannered observations can ruin your analysis. In H. T. Reis & C. M. Judd (Eds.). *Handbook of research methods in social and personality psychology* (pp. 393-411). Cambridge: Cambridge University Press.