

Data Cleaning Additional Resources

Websites

<http://www.childrensmercy.org/stats/category/DataManagement.asp>
<http://www.mathworks.com/products/statistics/description2.html>
http://www.spss.com/spss/data_mgmt.htm
http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/Missing.html
<http://www.spc.uchicago.edu/DATALIB/DLguides/Gdathand.html>
<http://www.cdc.gov/hiv/software/pubs/codebook.pdf>
http://ori.dhhs.gov/education/products/n_illinois_u/datamanagement/dotopic.html
<http://www.worldagroforestry.org/newwebsite/sites/rsu/datamanagement/documents/Session7/IssuesOnDataOwnership.asp>
http://research.wayne.edu/compliance/Data_Ownership_5_2_062.pdf
<http://www.mcmaster.ca/senate/academic/ownstwrk.htm>
<http://edweb6.educ.msu.edu/kin/BylawsPolicies/Research.pdf>
http://web.uvic.ca/psyc/grad/grad-rules/Appendix_K.htm
<http://ies.ed.gov/pubsearch/pubsinfo.asp?pubid=NCEE20090049>
<http://psychology.illinoisstate.edu/jccutti/138web/spss/spss3.html>
<http://pareonline.net/getvn.asp?v=9&n=6>
<http://phdmonkey.com/2011/07/15/what-is-a-data-codebook/>
<http://web.pdx.edu/~cgrd/codebk.htm>
http://www.tulane.edu/~panda2/Analysis2/submods/Code%20book/create_a_code_book.htm
http://dss.princeton.edu/online_help/analysis/codebook.htm
<http://toolkit.pellinstitute.org/evaluation-guide/analyze/enter-organize-clean-data/>
http://www.stepstoolkit.org/index.php?option=com_courses&task=ShowModule&Module=8&type=T&CatId=54&Itemid=142&lang=en
<http://www.lib.umich.edu/research-data-management-and-publishing-support/nsf-data-management-plans>
<http://www.meadinkent.co.uk/xlfreq.htm>
http://statistics-help-for-students.com/What_are_Z_scores.htm#.UIWjn67YFQg
http://en.wikipedia.org/wiki/Standard_score
<http://en.wikipedia.org/wiki/Outlier>
<http://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm>
<http://pareonline.net/getvn.asp?v=9&n=6>
http://online.stat.psu.edu/online/development/stat501/14outliers/02outlier_distinction.html
<http://asq.org/quality-progress/2010/02/statistics-roundtable/outlier-options.html>
<http://mvint.usbmed.edu.co:8002/ojs/index.php/web/article/viewFile/460/442>
http://en.wikipedia.org/wiki/Normality_test
<http://core.ecu.edu/psyc/wuenschk/stathelp/NormalityAssumption.htm>
<http://www.psychology.nottingham.ac.uk/staff/pal/stats/C82MST/C82MST%20Lecture%202a%20Notes.htm>

Youtube Videos

<http://www.youtube.com/watch?v=R6Cc5flsbsw> (*data codebook*)

http://www.youtube.com/watch?v=js1s_tDUQmE&feature=relmfu (*datasets*)

<http://www.youtube.com/watch?v=NZXF811CzC8&feature=relmfu> (*data*)

<http://www.youtube.com/watch?v=BnoYEcycg->

[HM&playnext=1&list=PL5DE86FDC53716BEE&feature=results_main](http://www.youtube.com/watch?v=BnoYEcycg-HM&playnext=1&list=PL5DE86FDC53716BEE&feature=results_main) (*spss for newbies; many videos by this person*)

<http://www.youtube.com/watch?v=hUY3skSprvo> (*data cleaning*)

<http://www.youtube.com/watch?v=QzdSV0z1DB0> (*data cleaning*)

<http://www.youtube.com/watch?v=5qhLDYr70MM&feature=channel&list=UL> (*data screening*) (1 of 10 videos)

<http://www.youtube.com/watch?v=BSbAiQu6Ekc&feature=channel&list=UL> (*exploring data*)

<http://www.youtube.com/watch?v=XIUdRVdT7iU> (*missing data*)

http://www.youtube.com/watch?v=dnm13d0ofek&playnext=1&list=PL5E292F04CD37E587&feature=results_main (*frequency analysis*)

<http://www.youtube.com/watch?v=iRsH61VHij0&feature=related> (*frequency analysis*)

<http://www.youtube.com/watch?v=outMas3NrQ8&feature=related> (*frequency analysis*)

<http://www.youtube.com/watch?v=WSfISmcNRFI&feature=related> (*outliers*)

<http://www.youtube.com/watch?v=3R-DUoSoV7I> (*creating composite variables*)

<http://www.youtube.com/watch?v=AHlgb-BFNdM> (*computing variables in SPSS*)

http://www.youtube.com/watch?v=okf2noki_tE (*z-scores*)

<http://www.youtube.com/watch?v=1CHwYvV5tVY> (*z-scores*)

http://www.youtube.com/watch?v=_Qrqd9jl9Io (*descriptive statistics*)

<http://www.youtube.com/watch?v=YjqThWATgh4> (*descriptive statistics*)

<http://www.youtube.com/watch?v=fg70wcEiw6k> (*outliers*)

<http://www.youtube.com/watch?v=sQkB-AlJgPI> (*normality*)

http://www.youtube.com/watch?v=9a9Lp-1n_ZU (*normality*)

<http://www.youtube.com/watch?v=4yhpYKzW98M> (*replacing missing data – part 1*)

<http://www.youtube.com/watch?v=xEkJxl6mmQ0> (*replacing missing data – part 2*)

<http://www.youtube.com/watch?v=xbp2E8FdrGI> (*linearity*)

Software

Free qualitative software: <http://www.umass.edu/qdap/>

Free statistical software: http://davidmlane.com/hyperstat/Statistical_analyses.html

Free SPSS like software: <http://www.gnu.org/software/pspp/>

Published Articles/Books

- Cody, R. P. (2008). *Cody's data cleaning techniques using SAS*. SAS Institute.
- Cole, J. C. (2008). *How to deal with missing data*. In Osborne, J. W. (Ed.) Best practices in quantitative methods (pp.214-238). Sage: Thousand Oaks, CA. doi: 10.4135/9781412995627.
- Dasu, T., & Johnson, T. (2003). *Exploratory data mining and data cleaning* (Vol. 479). Wiley-Interscience.
- DeCuir-Gunby, J.T., Marshall, P.L., & McCulloch, A.W. (2011). Developing and using a codebook for the analysis of interview data: An example from a professional development research project. *Field Methods*, 23(2) 136-155.
- Guillet, F., & Hamilton, H. J. (Eds.). (2007). *Quality measures in data mining* (Vol. 43). Springer.
- Han, J., Kamber, M., & Pei, J. (2006). *Data mining: concepts and techniques*. Morgan kaufmann.
- Hellerstein, J.M. (2008). *Quantitative data cleaning for large databases*. <http://db.cs.berkeley.edu/jmh>.
- Israel, G.D. (1992). *Phases of data analysis*. Program Evaluation and Organizational Development, IFAS, University of Florida. PEOD-1. October.
- Israel, G.D. (1992). *Elaborating program impacts through data analysis*. Program Evaluation and Organizational Development, IFAS, University of Florida. PEOD-3, September.
- Janert, P. K. (2010). *Data Analysis with Open Source Tools*. O'Reilly Media, Inc.
- Nimon, K. F. (2012). Statistical assumptions of substantive analyses across the general linear model: a mini-review. *Frontiers in Psychology*, 3.
- Osborne, J. W. (2008). *Best practices in data transformation: the overlooked effect of minimum values*. In Osborne, J. W. (Ed.) Best practices in quantitative methods (pp.197-204). Sage: Thousand Oaks, CA. doi: 10.4135/9781412995627.
- Osborne, J. W. (2012). *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data*. SAGE Publications, Incorporated.
- Osborne J. & Overbay, A. (2008). *Best Practices in data cleaning: how outliers and "fringeliers" can increase error rates and decrease the precision of your results*. In Osborne, J. W. (Ed.) Best practices in quantitative methods (pp.205-213). Thousand Oaks, CA: Sage Publications. doi: 10.4135/9781412995627.
- Parke, C. S. (2012). *Essential First Steps to Data Analysis: Scenario-based Examples Using SPSS*. SAGE Publications, Incorporated.
- Tabachnick, B. G., & Fidell, L. S. (2013). *Using Multivariate Statistics*, 6th ed. Upper Saddle River, NJ: Pearson.
- Van den Broeck, J., Cunningham, S. A., Eeckels, R. & Herbst, K. (2005). Data cleaning: detecting, diagnosing, and editing data abnormalities. *PLoS Medicine*, 2, (10), 0966-0970. doi: 10.1371/journal.pmed.0020267.

- Wilkinson, L., & APA Task Force on Statistical Inference. (1999). *Statistical methods in psychology journals: Guidelines and explanations*. *American Psychologist*, 54, 594-604. [reprint available through the APA Home Page: <http://www.apa.org/journals/amp/amp548594.html>]
- Wuensch, K. L. (2005). *Kurtosis*. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1028 - 1029). Chichester, UK: Wiley.
- Wuensch, K. L. (2005). *Skewness*. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1855 - 1856). Chichester, UK: Wiley

