# GCD Theory week 2 – Getting and Cleaning Data

**Objectives**

The student:

* Knows the 5 measures of data quality
* Is able to assess data on these measures
* Is able to develop strategies to cleanup data

**Preparation**

Watch the following videos from Udacity Data Wrangling with MongoDB. The code ‘Lesson 1.3’ means within ‘Lesson 1’, the 3rd part of the lesson.

|  |  |  |  |
| --- | --- | --- | --- |
| **Title** | **Duration** | **Udacity Lesson** | **Week** |
| Assessing the Quality of Data (part 1) | 0:14 | 1.3 | 2 |
| Assessing the Quality of Data (part 2) | 2:10 | 1.4 | 2 |
| What is Data Cleaning? | 0:50 | 3.1 | 2 |
| Sources of Dirty Data | 1:09 | 3.4 | 2 |
| Measuring Data Quality | 1:31 | 3.5 | 2 |
| Difficulty of Quality Metrics | quiz | 3.6 | 2 |
| A Blueprint for Cleaning | 1:46 | 3.7 | 2 |
| Auditing Validity | 1:28 | 3.9 | 2 |
| Auditing Accuracy | 1:05 | 3.13 | 2 |
| Auditing Accuracy 2 | 4:06 | 3.14 | 2 |
| Auditing Completeness | 2:35 | 3.15 | 2 |
| Auditing Consistency | 1:05 | 3.16 | 2 |
| Consistency | 2:03 | 3.17 | 2 |
| Auditing Uniformity | 5:07 | 3.18 | 2 |
| More about Correcting Data | 1:58 | 3.19 | 2 |

Missing data: watch this presentation (detailed, but slow) (16min): <https://www.youtube.com/watch?v=XlUdRVdT7iU>

If you are interested, or need extra information, read pages 2 and further.

**In class activities**

In the videos the following 5 measures of Data Quality are mentioned:

* Validity
* Accuracy
* Completeness
* Consistency
* Uniformity

In class we will **create** a dataset and **analyse** the quality of data on the 5 measures and make a **strategy** how to cleanup the data and handle missing data.

## Theory Missing Data

Video: detailed overview, but slow, 16m: <https://www.youtube.com/watch?v=XlUdRVdT7iU>

2. Missing data mechanisms

There are different assumptions about missing data mechanisms:

1. Missing completely at random (MCAR): Suppose variable Y has some missing values. We will say that these values are MCAR if the probability of missing data on Y is unrelated to the value of Y itself or to the values of any other variable in the data set. However, it does allow for the possibility that “missingness” on Y is related to the “missingness” on some other variable X. (Briggs et al., 2003) (Allison, 2001)

Example: We want to assess which are the main determinants of income (such as age). The MCAR assumption would be violated if people who did not report their income were, on average, younger than people who reported it. This can be tested by dividing the sample into those who did and did not report their income, and then testing a difference in mean age. If we fail to reject the null hypothesis, then we can conclude that the MCAR is mostly fulfilled (there could still be some relationship between missingness of Y and the values of Y).

1. Missing at random (MAR)-a weaker assumption than MCAR: The probability of missing data on Y is unrelated to the value of Y after controlling for other variables in the analysis (say X). Formally: P(Y missing|Y,X) = P(Y missing|X) (Allison, 2001).  
   Example: The MAR assumption would be satisfied if the probability of missing data on income depended on a person’s age, but within age group the probability of missing income was unrelated to income. However, this cannot be tested because we do not know the values of the missing data, thus, we cannot compare the values of those with and without missing data to see if they systematically differ on that variable.
2. Not missing at random (NMAR): Missing values do depend on unobserved values.  
   Example: The NMAR assumption would be fulfilled if people with high income are less likely to report their income.

If MAR assumption is fulfilled: The missing data mechanism is said to be ignorable, which basically means that there is no need to model the missing data mechanism as part of the estimation process. These are the method this report will cover.  
If MAR assumption is not fulfilled: The missing data mechanism is said to be nonignorable and, thus, it must be modeled to get good estimates of the parameters of interest. This requires a very good understanding of the missing data process.

[Source: <http://www.bu.edu/sph/files/2014/05/Marina-tech-report.pdf> ]

4. Methods for handling missing data

4.1.

4.1.1.  Listwise deletion (or complete case analysis): If a case has missing data for any of the variables, then simply exclude that case from the analysis. It is usually the default in statistical packages. (Briggs et al.,2003).

Advantages: It can be used with any kind of statistical analysis and no special computational methods are required.  
Limitations: It can exclude a large fraction of the original sample. For example, suppose a data set with 1,000 people and 20 variables. Each of the variables has missing data on 5% of the cases, then, you could expect to have complete data for only about 360 individuals, discarding the other 640.

It works well when the data are missing completely at random (MCAR), which rarely happens in reality (Nakai & Weiming, 2011).

4.1.2.  Imputation methods: Substitute each missing value for a reasonable guess, and then carry out the analysis as if there were not missing values.

Conventional methods

There are two main imputation techniques:  
▪ Marginal mean imputation: Compute the mean of X using the non-missing values and use it to impute missing values of X.  
Limitations: It leads to biased estimates of variances and covariances and, generally, it should be avoided.

▪ Conditional mean imputation: Suppose we are estimating a regression model with multiple independent variables. One of them, X, has missing values. We select those cases with complete information and regress X on all the other independent variables. Then, we use the estimated equation to predict X for those cases it is missing.

If the data are MCAR, least-squares coefficients are consistent (i.e. unbiased as the sample size increases) but they are not fully efficient (remember, efficiency is a measure of the optimality of an estimator. Essentially, a more efficient estimator, experiment or test needs fewer samples than a less efficient one to achieve a given performance). Estimating the model using weighted least squares or generalized least squares leads to better results (Graham, 2009) (Allison, 2001) and (Briggs et al., 2003).

Limitations of imputation techniques in general: They lead to an underestimation of standard errors and, thus, overestimation of test statistics. The main reason is that the imputed values are completely determined by a model applied to the observed data, in other words, they contain no error (Allison, 2001).

Statistics has developed two main new approaches to handle missing data that offer substantial improvement over conventional methods: Multiple Imputation and Maximum Likelihood.

[Source: <http://www.bu.edu/sph/files/2014/05/Marina-tech-report.pdf> ]

Because we often gather data on multiple related variables, we often know (or can estimate) a good deal about the missing values. Aside from examining missingness as an outcome itself (which I strongly recommend), modern computing affords us the opportunity to fill in many of the gaps with high-quality data. This is not merely “making up data” as some early, misinformed researchers claimed. Rather, as my examples show, the act of estimating values and retaining cases in your analyses most often leads to more replicable findings as they are generally closer to the actual population values than analyses that discards those with missing data (or worse, substitutes means for the missing values). Thus, using best practices in handling missing data makes the results a better estimate of the population you are interested in. And it is surprisingly easy to do, once you know how.

Thus, it is my belief that best practices in handling missing data include the following.

* First, do no harm. Use best practices and careful methodology to minimize missingness. here is no substitute or complete some careful forethought can often save a good deal of frustration in the data analysis phase of research.
* Be transparent. Report any incidences of missing data (rates, by variable, and reasons for missingness, if possible). This can be important information to reviewers and consumers of your research and is the first step in thinking about how to effectively deal with missingness in your analyses.
* Explicitly discuss whether data are missing at random (i.e., if there are differences between individuals with incomplete and complete data). Using analyses similar to those modeled in this chapter, you can give yourself and the reader a good sense of why data might be missing and whether it is at random. That allows you, and your audience, to think carefully about whether missingness may have introduced bias into the results. I would advocate that all authors report this information in the methods section of formal research reports.
* Discuss how you as a researcher have dealt with the issue of incomplete data and the results of your intervention. A clear statement concerning this issue is simple to add to a manuscript, and it can be valuable for future consumers as they interpret your work. Be specific—if you used imputation, how was it done, and what were the results? If you deleted the data (complete case analysis) justify why.

Finally, as I mentioned in Chapter 1, I would advocate that all authors report this information in the methods section of formal research reports and that all journals and editors and conferences mandate reporting of this type. If no data is missing, state that clearly so consumers and reviewers have that important information as well.

[Source: [https://us.sagepub.com/sites/default/files/upm-binaries/45664\_6.pdf p.131](https://us.sagepub.com/sites/default/files/upm-binaries/45664_6.pdf%20p.131)]

## Practical information

Scikit-Learn: <http://scikit-learn.org/stable/modules/preprocessing.html#imputation-of-missing-values>