COM-3590: Data Cleaning and Transformation

“When statistics are not based on strictly accurate calculations, they mislead instead of guide. The mind easily lets itself be taken in by the false appearance of exactitude which statistics retain in their mistakes, and confidently adopts errors clothed in the form of mathematical truth.” —Alexis de Tocqueville, Democracy in America

Learning Objectives

Course Level Learning Outcomes

Collect, clean, and prepare data.

Evaluate data in terms of source, volume, frequency, and flow.

Understand data and gather maximum information out of it with Pattern Detection using Data Mining, Data Mapping and Clustering.

Week 1.

**Course overview**. Garbage in, garbage out: how dirty data can impact analysis.

**Objectives:**

* Understand the importance of students to clean, organize, enhance data before inserting it into a database or merging it with other data files.
  + *Before a data item ends up in a database, it typically passes through a number of steps involving both human interaction and computation. Data errors can creep in at every step of the process from initial data acquisition to archival storage. An understanding of the sources of data errors can be useful both in designing data collection and curation techniques that mitigate the introduction of errors, and in developing appropriate post-hoc data cleaning techniques to detect and ameliorate errors.*
* sources of errors in databases
* Data Types
  + *Quantitative Data, Categorical Data, Postal Addresses, Identifiers (or keys)*
* Cleaning Techniques

**To know garbage in, garbage out, data collection, curation, type 1 & type 2**

**To understand that data errors can creep in every step of the process.**

**To be able to identify errors in data and artifacts.**

**Evidence: Create a profile/scenario where a company incorrectly drew conclusions or wasted resources because they did not pay attention to data. Students vocalize process through which they analyzed data. Write step-by-step data.**

**Questions:**

* What are the main sources of errors?
  + *Data entry errors, Measurement errors, Distillation errors, Data integration errors*
* What are the approaches for Improving Data Quality?
  + *Data entry interface design, Organizational management, Automated data auditing and cleaning, Exploratory data analysis and cleaning*

**Errors vs. artifacts.**

Type 1 error. Mistakenly concluding an effect is real (when it is due to chance).

Type 2 error. Mistakenly concluding an effect is due to chance (when it is real).

Sources of errors in data and their telltale signs in data sets.

Questions. How can we prevent wrong conclusions from data

**Week 2.**

**Exploratory Data Analysis (EDA).** Examine the data for distribution, outliers and anomalies to direct specific testing of your hypothesis.

**Objectives:**

* Maximize insight into the database/understand the database structure;
* Visualize potential relationships (direction and magnitude) between exposure and outcome variables;
* Detect outliers and anomalies (values that are significantly different from the other observations);
* Develop parsimonious models (a predictive or explanatory model that performs with as few exposure variables as possible) or preliminary selection of appropriate models;
* Extract and create clinically relevant variables.

**Questions:**

* Why is EDA important during the initial exploration of a dataset?
* What are the most essential tools of graphical and non-graphical EDA?

**Visualization Tools.** Synthesis Relationships between data

**Objectives:**

* Develop exploratory data analysis and visualization tools using Python and Jupyter notebooks
* Apply design principles for a variety of statistical graphics and visualizations including scatterplots, line charts, histograms, and choropleth maps
* Combine exploratory queries, graphics, and interaction to develop functional tools for exploratory data analysis and visualization

**Questions:**

**Regular Expressions** (RE): A pattern that specifies a set of character strings. REs are most often used to find sequences of characters in strings. Think: search and replace

**Objectives:**

* Review regex pattern syntax
* Be able to perform both simple and more complex searches and replacements
* Know how to reference matches by capturing groups

**Questions:**

How do you find words that match a particular pattern?

**Open Refine**. Tool for cleaning large datasets

**Objectives:**

* Introduce Open Refine as a powerful data-cleaning tool.
* Encourage dataset exploration; look at the data with the visualization tools in Open Refine.

**Questions:**

How would you keep track of all steps applied to your data?

How can you speed up repetitive tasks by replaying previous actions on multiple datasets?

**Week 3**

**Outliers.** According to Hawkins (1980), “An outlier is an observation that deviates so much from

other observations as to arouse suspicion that it was generated by a different mechanism”.

**Univariate outlier detection**. (only one variable, exposure or outcome) A univariate outlier is a data point that consists of an extreme value on one variable.

Robust Statistics and Estimators

Questions:

How do I detect outliers?

<http://www.psychwiki.com/wiki/Detecting_Outliers_-_Univariate>

**Week 4**

Multivariate outlier detection. (several exposure variables alone or with an outcome variable) A multivariate outlier is a combination of unusual scores on at least two variables

Robust Multivariate Estimation.

Questions: Why outliers exists?

*(Outliers exist for four reasons. Incorrect data entry can cause data to contain extreme cases. A second reason for outliers can be failure to indicate codes for missing values in a dataset. Another possibility is that the case did not come from the intended sample. And finally, the distribution of the sample for specific variables may have a more extreme distribution than normal.)*

How can Multivariate outliers be identified?

Multivariate outliers can be identified with the use of Mahalanobis distance, which is the distance of a data point from the calculated centroid of the other cases where the centroid is calculated as the intersection of the mean of the variables being assessed. Each point is recognized as an X, Y combination and multivariate outliers lie a given distance from the other cases. The distances are interpreted using a p < .001 and the corresponding χ2 value with the degrees of freedom equal to the number of variables. Multivariate outliers can also be recognized using leverage, discrepancy, and influence. Leverage is related to Mahalanobis distance but is measured on a different scale so that the χ2 distribution does not apply. Large scores indicate the case if further out however may still lie on the same line. Discrepancy assesses the extent that the case is in line with the other cases. Influence is determined by leverage and discrepancy and assesses changes in coefficients when cases are removed. Cases > 1.00 are likely to be considered outliers.

**Week 6**

**Resampling Techniques.** Repeatedly sample values from observed data, with a general goal of assessing random variability in a statistic.

**Objectives:**

**Questions:**

**Frequency outliers.**

**Week 7**

**Data Sampling:**

* Observation unit: An object on which a measurement is taken.
* Target population: The complete collection of observations we want to study.
* Sample: A subset of a population.
* Sampled population: The collection of all possible observation units that might have been chosen in a sample; the population from which the sample was taken

**Sample and selection bias**.

* Selection bias occurs when some part of the target population is not in the sampled population, or, more generally, when some population units are sampled at a different rate than intended by the investigator.

**Objectives:**

* Differentiate between a census and a survey or sample.
* Distinguish between sampling error and bias.
* Identify and name potential sources of bias from both real and hypothetical sampling situations.

**Questions:**

* Why sample at all?
* What are the main justifications for using sampling?
* How is the margin of error for a survey calculated?
* What are the effects of sample size on sampling error?
* Is the plural of census censuses, or censi?

**Confidence intervals**.

**Objective:**

* Calculate the point estimate of a sample to estimate a population proportion.
* Construct a confidence interval for a population proportion based on a sample population.
* Calculate the margin of error for proportions as a function of sample proportion and size.
* Understand the logic of confidence intervals as well as the meaning of confidence level and confidence intervals.

**Distributions**: normal, t, binomial, Poisson.

**Week 8**

**Deduplication. Case study: sales data, accounting changes, and the surprise jump in revenue**

**Week 9**

Data Transformation: Restructuring, Enriching.

**Week 10**

Univariate entity resolution (a.k.a. string matching)

**Week 11**

Schema Matching and Mapping

**Week 12**

Multivariate entity resolution (a.k.a. data matching)

**Week 13**

Review

Background correction and normalization (April 28-May 4)

Genomics data is usually affected by systematic bias and unwanted variability. Without proper normalization, it can often be impossible to compare across samples. This week we will take a look at statistical models in the context of genomics data. In particular we will explore how to describe and adjust for background noise and achieve normalization with data.

Objectives:

Describe the typical distribution and joint distribution of microarray data.

Distinguish between several approaches to background adjustments.

Identify the impact of background noise on microarray data.

Explain why normalization is necessary for proper interpretation of genomics data.

Describe the different normalization approaches for various microarray technologies.

Enumerate special cases for which current normalization techniques are not appropriate.

Batch effects (May 12-May 18)

In this session, we will explore some useful statistical ideas for accounting for batch effects as well as looking at specific solutions for batch effects.

Objectives:

Identify and remove batch effects in microarray data.

Describe the usefulness of PCA in finding batch effects in high dimensional data.

Sections:

Statistical solutions to batch effects

Applying batch effects solutions

References:

Regular Expressions: <https://www.ibm.com/developerworks/library/j-perry-regular-expressions/index.html>

Categorical Measures

A commonly used expression for nominal and ordinal measues

Frequency Analysis

A count of the number of cases that fall into each category when the categories are based on one variable

Outlier

An observation so different in magnitude from the rest of the observations that the analyst chooses to treat it as a special case

Histogram

A form of bar chart on which the values of the variable are placed along the x-axis and the absolute frequency or relative frequency of occurrence of the values is indicated along the y-axis

Confidence Interval

A projection of the range within which a population parameter will lie at a given level of confidence based on a statistic obtained from a probabilistic sample

Continuous Measures

A commonly used expression for interval and ratio measures

Descriptive Statistics

Stats that describe the distribution of responses on a variable. The most commonly used descriptive stats are the mean and std deviation

- Variation in distribution: range, variance, & std dev

- Central tendencies: mean, median, mode

- Shape of the Distribution: skewness, kurtosis

Sample Mean

Arithmetic average value of the response on a variable

Sample Std Deviation

A measure of the variation of responses on a variable

the std dev is the square root of the calculated variance on a variable

Median Split

A technique for converting a continuous measure into a categorical measure with 2 approx. equalsized groups. The groups are formed by "splitting" the continuous measure at the median value

Cumulative Percent Breakdown

A technique for converting a continuous measure into a categorical measure. the categories are formed based on the cumulative percentages obtained in a frequency analysis

Two-Box Technique

A technique for converting an interval-level rating scale into a categorical measure usually used for presentation purposes

The percentage of respondents choosing one of the top two positions on a rating scale is reported

Hypothesis

Unproven propositions about some phenomenon of interest

Null Hypothesis

The hypothesis that a proposed result is not true for the population. Researchers typically attempt to reject the null hypothesis in favor of some alternative hypothesis

Alternative Hypothesis

Hypothesis that a proposed result is true for the population

Significance Level

The acceptable level of Type I error selected by the researcher, usually set at .05

Type I Error

The probability of rejecting the null hypothesis when it is actually true for the population

P-Value

The probability of obtaining a given result if in fact the null hypothesis were true in the population. A result is regarded as statistically significant if the p-value is less than the chosen significance level of the test

Chi-Square Goodness-of-Fit Test

A statistical test to determine whether some observed pattern of frequencies corresponds to an expected pattern

Distinguish b/w Univariate and Multivaraite Analyses

Univaiate analyses are conducted on individual variable; multivariate analyses involve multiple variables

Describe Frequency Analysis

A univariate technique that involves counting the # of responses that fall into various response categories

Various Ways in which Frequency Analysis Can be Used

(1) to communicate the results of the story

(2) to determine the degree of item nonresponse

(3) to locate blunders

(4) to locate outliers

(5) to determine the empirical distribution of the variable in question

Discuss confidence intervals for proportions

the confidence interval is the range within the true proportion in the population will fall given a level of confidence. the confidence interval is equal to the sample proportion plus or minus estimated sampling error

Describe commonly used descriptive statistics.

the most commonly used descriptive stats for continuous measures are the mean & std dev

Discuss confidence intervals for means

The confidence interval is the range within the true mean value for the population will fall with a given level of confidence

the confidence interval is = to the sample mean plus of minus estimated sampling error

Steps involved in hypothesis testing

(1) Specify null and alterative hypothesis

(2) Choose the appropriate test statistic

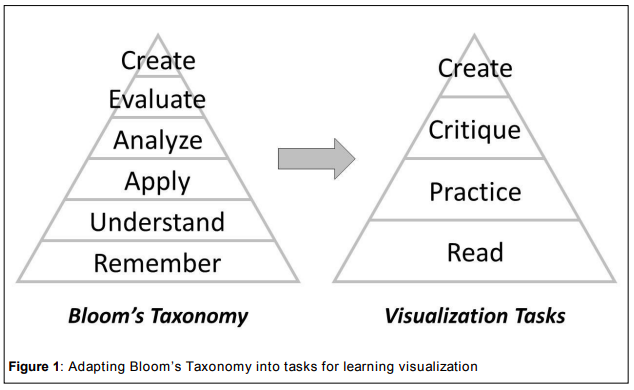
(3) Specify significance level for the test

(4) Compute value of test statistic based on sample data

(5) Determine probability of test stat given true null hypothesis

(6) Compare obtained probability with specified significance level

At the base of the Taxonomy, the basic cognitive processes of “Remember” and “Understand” become “Read,” a task to learn about the history and theory of data visualization. The next Bloom’s level, “Apply,” becomes “Practice” and entails going through the mechanics of producing different charts in a chosen software program. The higher Bloom’s levels “Analyze” and “Evaluate” become “Critique,” which requires the librarian to evaluate existing visualizations using lessons learned through Reading and Practice. Finally, the Bloom’s “Create” category becomes the “Create” task, in which the librarian brings their new knowledge and skills together to design effective visualizations. Working through each task raises skills and fluency in data visualization, ultimately positioning the librarian to confidently develop services around this competency.



Source: https://escholarship.umassmed.edu/cgi/viewcontent.cgi?article=1116&context=jeslib

## Motivations for the Open Refine Lesson

* It’s really important that you know what you did. More journals/grants/etc. are also making it important for them to know what you did. You can **capture** all steps done to your data in Open Refine to the raw file and share them with your publication as supplemental material.
* All steps are easily **reversed** in Refine.
* You must save your work to a new file; Refine does not modify your original dataset.
* Data is often very **messy** - and this tool saves a lot of **cleaning** headaches.
* Data cleaning steps often need repeating with multiple files. Refine is perfect for speeding up **repetitive tasks**.
* Some concepts like clustering algorithms are quite complex, but Refine makes it easy to introduce them, use them, and show their power.