CPSC448 Report (May – August 2018)

EDUCE = Experiential Data Science for Undergraduate Cross-disciplinary Education

Project Website with Apps = https://educe-ubc.github.io/

Project Components:

- Static website for the EDUCE Project.
- Course Compiler App for the EDUCE Project.
- Data Manipulator App for the EDUCE Project that includes user generated integration testing (CPSC) and various methods of normalizing data (MICB/STAT).

Purpose of Project:

There are various objectives for this project:

CPSC Objective:
- Use this project as a testing ground as an integration test generator based on user inputs and discuss the benefits/drawbacks for this method of generating tests.

MICB Objective:
- Design GUI/App that will be useful for generating teaching materials provided in the EDUCE Project for data science instructors (novice/veteran) so that they will spend less time compiling course components and more time trying to understand the subject (novice) and/or assisting students with data science (veteran).

CPSC/MICB/STAT(?) Objective:
- Implement/Discuss the various benefits/drawbacks of methods of normalizing data.
- Design an App that helps with normalizing the inputted data and give the user a table of normalized data in text file format.
- Overall: Combining the use of interdisciplinary knowledge (mainly CPSC + MICB) in this project to help others in learning about both subjects via EDUCE.

What I learned (Summary):

From this project, I learned a variety of things from the CPSC, MICB, and STAT prospective:

For the CPSC component, I learned about the various benefits and drawbacks of user generated integration testing. In addition, I have also covered other parts of CPSC related topics that were not in my project proposal: Improvements of App efficiency and user readability of source code via refactoring and designing generic source code (Course Compiler) for reuse by veteran users.

For the MICB/STAT component, I learned about the benefits and drawbacks of various methods of normalizing data via Percent Relative Abundance AKA Relative Species Abundance (PRA), Random Subsampling AKA rarefaction (RS), and other potential methods of normalization not included in the App due to limitations such as Multiple Imputation (MI) and Variance Stabilizing Transformation (VST). I have also learned about designing Apps that are more user-friendly to novice users with the assumption that the majority of the users will have no coding experiences.

What I could improve on:

From the feedbacks given to me in the project, I could have approached the project in a better way by formulating a more detailed work plan (i.e. workflow) to work more effectively. I could also put more time into researching the two other methods of normalization/data transformation in this project (MI + VST).

User-Generated Integration Testing

Integration testing involves the testing of multiple components of the application to ensure that the components are working correctly together [1]. User-generated integration testing involves the generation of test cases from user's inputs that helps test the application under multiple conditions such as different file types, file contents, input types, etc.

For this project, the "Data Manipulator" App was used as the testing ground for User-Generated Integration Testing. The Big Bang Approach was used for the integration testing where all the components are integrated together at once and then tested [2]. To ensure that the correct validation data was saved, error catching mechanisms were implemented to limit invalid data [3].

Potential Benefits:

- Arbitrary test cases can be generated based on user inputs and files (assuming the users are using the correct data format/contents and error catching mechanisms are working).
- Reduced time in generating test cases, which can be used to improve the application in terms of efficiency and other bugfixes.
- Users can help identify bugs/issues faster via their usage of the App with their files/inputs since they may understand their data better (100 MICB Users with MICB data vs. 1 CPSC Developer that may or may not understand the MICB data being processed).

Potential Drawbacks:

- Significant resources are required in the long run such as storage space for user files/data. Can be resolved by limiting the number of user files/data saved.
- Some components may not be tested due to limitations:
    - Data/File Type = Since technology is constantly evolving, if the data/file type is changed then the integration testing codes may need to be modified.
    - File Size = Larger files may take longer to process during integration testing.
    - File Contents = Some contents may not be suitable to specific tests, such as normalizing data that are all 0 in this case (cannot divide by 0).
- Users may potentially choose not to share their data for integration testing purposes due to various reasons such as confidentiality of contents. Can be resolved by the owner(s) or developer(s) via submitting some initial test cases to be used in integration testing.
- If an error was made in designing the user-generated integration testing code, then potential errors may go unnoticed.
- Malicious users may flood the storage with duplicate data, which may reduce the effectiveness of user-generated integration testing.

Data Normalization

Data normalization is the process of transforming/standardizing data to a common scale for comparison [4]. This is especially useful for microbial ecology since the data may come from different samples and a way to compare the trends of data on a standardized scale will be useful [5]. The objective of this part of the project is to find potential ways of minimizing the effects of non-parametric (data that does not follow a normal distribution) and skewed data.

Methods of Data Normalization

Percent Relative Abundance (AKA Relative Species Abundance)

Percent Relative Abundance is a technique that transforms the data into percentages, which is then used to compare how the data is relative to one another. Also known as Relative Species Abundance in microbial ecology, it is a measure of how common a species is relative to other species in a defined location [6].

This technique first identifies the total number of observations to be represented in each group, then each data point is then divided by the total number of observations calculated earlier. The sum of all the percentages in each group should be 100% [7].

Potential Benefits:

- Can easily conceptualize data without overthinking (Dealing with percentages vs. Dealing with actual numbers without context) [8].
- Easy to transform data:

$$\frac{\text{Number of observations for each data point}}{\text{Total number of observations for each group of data points}}$$

Potential Drawbacks:

- Difficult to comprehend complex datasets [8]. Can be resolved by dividing large datasets into interval classes for easier visualization.

Random Subsampling (AKA Rarefaction)

Random Subsampling is technique that splits the data into subsets [9]. Also known as rarefaction in microbial ecology, it is a technique used to determine species richness of samples that differ in area, volume, or sampling efforts [10].

Potential Benefits:

- Subsetting of samples can be repeated an indefinite number of times.
- Rarefaction compares observed richness among samples for a given level of sampling effort and does not attempt to estimate true richness of community [11]:

$$\frac{\text{Number of observations for each subsampled data point}}{\text{Total number of observations for each group of subsampled data points}}$$

Potential Drawbacks:

- Rarefied counts remain over-dispersed relative to Poisson model (implies an increase in Type I error). More difficult to estimate overdispersion after rarefying due to lost information [12].
- Rarefied counts represent only a small fraction of original data (can lead to increased chance of Type II error). Often caused by samples being discarded and poorly distinguishable samples due to discarded fraction of original library [12].
- Random step in rarefying is unnecessary and adds artificial uncertainty [12].
- Many assumptions are required if rarefaction is used for comparing samples: Sufficient sampling, comparable sampling methods, taxonomic similarity, closed communities of discrete individuals, random placement, and independent random sampling [13, 14].

Multiple Imputation

Multiple Imputation is a statistical technique that is useful for analyzing incomplete or missing data via 3 steps [15, 16]:

1) Imputation: Missing entries of datasets are filled in *m* times, which can contain different values for each missing entry (*m* Complete datasets).
2) Analysis: The *m* completed datasets are then analyzed (*m* Analyses).
3) Pooling: The *m* analysis results are then pooled together into a final result.

The MICE Package was used in R with Predictive Mean Matching (PMM) as the imputation method (a list of imputation methods can also be found here) [17].

Potential Benefits:

- When Multiple Imputation is used correctly [18, 19, 20, 21, 22]:
    - Reduced bias: By imputing the missing values with imputation methods, analysis can be done on a "complete" dataset.
    - Improve validity: Valid inferences are obtained from the averaging of the distribution of missing data given the observed data.
    - Increase precision: Many missing data, which can potentially happen in microbial ecology due to the potential of excess zeros being missing data, may lead to the exclusion of a substantial proportion of the original sample, leading to a decrease in precision during analysis.
    - Result in robust statistics (resistant to outliers): Imputed values often uses the mean, median, or other statistic to impute the missing values.
- Useful for analyzing qPCR data, which often have moderate levels of missing values [20].

Potential Drawbacks:

- Must make a few required statistical assumptions [23]:
    - Is the data is missing at random?
    - Is this a multivariate normal distribution? (Needed for some of the modeling methods)
- Need to transform variables to approximate normal distribution before running imputation procedure (only if data isn't normally distributed) [18].
- Incorrect model choices or exclusion of vital data points may lead to more bias [18].

Variance Stabilizing Transformation

Variance Stabilizing Transformation is a technique that uses a function *f* to apply values to *x* in a dataset to create new values $y = f(x)$ such that the variability of values *y* is not related to their mean value (or has a constant variance) [24].

Potential Benefits:

- Usage of DESeq2 overcomes the issue of strong variance of logarithmic fold change (LFC) estimates for genes with low read count in analysis of high-throughput sequencing data [25]:
    - Shrinks LFC estimates toward zero when shrinkage is stronger when available information or gene is low, which could be due to low counts, high dispersion, or few degrees of freedom.
    - Prevents the spreading apart of data for genes with low counts, where random noise is likely to dominate any biologically meaningful signal (stabilizes variances).
- DeSeq2's heuristics can deal with flagged outliers in two possible ways to help avoid Type I errors [25]:
    - Outliers in conditions with six or fewer replicates cause whole gene to be flagged and removed from subsequent analysis (includes P value adjustment for multiple testing).
    - Outliers in conditions with seven or more replicates replaces the outlier counts with imputed value (trimmed mean over all samples), scaled by size factor, and then re-estimates dispersion, LFCs, and P values for these genes.
- Can consistently perform over large range of data types and is applicable for small studies with few replicates or large observational studies [25].

Potential Drawbacks:

- Many rare species are ignored completely due to the negative values from log-like transformations being set to 0. Since DESeq was developed mainly for use with Euclidean metrics (negative results are not a problem), this may yield misleading results for ecologically useful non-Euclidean measures [26].
- Many not be appropriate for highly diverse microbial environments since DESeq assumes that most microbes are not differentially abundance and for the few that are, there is an approximately balanced amount of increased/decreased abundance [26].

Acknowledgements

I would like to thank Dr. Kim Dill-McFarland (MICB), Dr. Steven Hallam (MICB), and Dr. Reid Holmes (CPSC) for supervising this directed study project. I would also like to thank Dr. Michael Gelbart (CPSC) and Ashley Arnold (MICB) for their contributions the multiple imputation and variance stabilizing sections, respectively.

References

[1] Holmes R. Software Testing. https://github.com/ubccpsc/310/blob/2017jan/readings/Testing.md (2018).

[2] Anonymous. Big Bang Integration Testing. http://www.professionalqa.com/big-bang-integration-testing (2018).

[3] Easterbrook S. Testing Strategies. http://www.cs.toronto.edu/~sme/CSC302/notes/17-Testing2.pdf (2018).

[4] Borgatti S. http://www.analytictech.com/ba762/handouts/normalization.htm (2018).

[5] Daniel Aguirre de Cárcer, Denman SE, McSweeney C, Morrison M. Evaluation of Subsampling-Based Normalization Strategies for Tagged High-Throughput Sequencing Data Sets from Gut Microbiomes. *Applied and Environmental Microbiology*. 2011; 77: 8795-8798.

[6] Socratic. How do species richness and relative abundance of species affect species diversity? https://socratic.org/questions/how-do-species-richness-and-relative-abundance-of-species-affect-species-diversi (2018).

[7] Percentage Frequency Distribution. *Encyclopedia of Survey Research Methods*. 2008.

[8] Reid A. Advantages & Disadvantages of a Frequency Table. https://sciencing.com/advantages-disadvantages-frequency-table-12000027.html (2018).

[9] Dieterle F. Random Subsampling. http://www.frank-dieterle.com/phd/2_4_3.html (2018).

[10] Chiarucci A, Bacaro G, Rocchini D, Ricotta C, Palmer MW, Scheiner SM. Spatially constrained rarefaction: incorporating the autocorrelated structure of biological communities into sample-based rarefaction. *Community Ecology*. 2009; 10: 209-214.

[11] Hughes JB, Hellmann JJ. The application of rarefaction techniques to molecular inventories of microbial diversity. In: Vol 397. United States: Elsevier Science & Technology; 2005: 292-308.

[12] McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Computational Biology*. 2014; 2013; 10: e1003531.

[13] Gotelli, NJ, Colwell RK. Estimating species richness. *Frontiers in Measuring Biodiversity*. 2011; 12: 39-54.

[14] Tipper JC. Rarefaction and Rarefiction; The Use and Abuse of a Method in Paleoecology. Paleobiology. 1979; 5: 423-434.

[15] van Buuren S. Multiple Imputation. http://www.stefvanbuuren.nl/mi/mi.html (2018).

[16] Maldonado, A. D.; Aguilera, P. A.; and Salmeron, A. An Experimental Comparison of Methods to Handle Missing Values in Environmental Datasets. *International Congress on Environmental Modelling and Software*. 2016: 3.

[17] van Buuren S. MICE. https://www.rdocumentation.org/packages/mice/versions/2.25/topics/mice (2018).

[18] Anonymous. Statistics How To. http://www.statisticshowto.com/multiple-imputation/ (2018).

[19] Sterne JAC, White IR, Carlin JB, *et al*. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009; 338: 157-160.

[20] Kanwar N, Scott HM, Norby B, *et al*. Impact of treatment strategies on cephalosporin and tetracycline resistance gene quantities in the bovine fecal metagenome. *Scientific Reports*. 2014; 2015; 4: 5100.

[21] Xu L, Paterson AD, Turpin W, Xu W. Assessment and Selection of Competing Models for Zero-Inflated Microbiome Data. *PLoS One*. 2015; 10: e0129606.

[22] Kaul A, Mandal S, Davidov O, Peddada SD. Analysis of Microbiome Data in the Presence of Excess Zeros. *Frontiers in Microbiology*. 2017; 8: 2114.

[23] Quora. How do you handle missing data (statistics)? What imputation techniques do you recommend or follow? https://www.quora.com/How-do-you-handle-missing-data-statistics-What-imputation-techniques-do-you-recommend-or-follow (2018).

[24] NC State University. Nonlinear Statistical Models for Univariate and Multivariate Response. https://www.stat.ncsu.edu/people/bloomfield/courses/ST762/slides/MD-02-2.pdf (2018).

[25] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*. 2014; 15: 550-550.

[26] Weiss S, Xu ZZ, Peddada S, *et al*. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*. 2017; 5: 27.