SESUG Paper 207-2023

CAMIS: Comparing Analysis Method Implementations in Software

Brian Varney, Experis

ABSTRACT

Several discrepancies have been discovered in statistical analysis results between different programming languages, even in fully qualified statistical computing environments. Subtle differences exist between the fundamental approaches implemented by each language, yielding differences in results, which are each correct in their own right. The fact that these differences exist causes unease on the behalf of sponsor companies when submitting to a regulatory agency, as it is uncertain if the agency will view these differences as problematic. In its Statistical Software Clarifying Statement, the US Food and Drug Administration (FDA) states that it "FDA does not require use of any specific software for statistical analyses" and that "the computer software used for data management and statistical analysis should be reliable." Observing differences across languages can reduce the analyst's confidence in reliability and, by understanding the source of any discrepancies, one can reinstate confidence in reliability.

The goal of this project is to demystify conflict when doing QC and to help ease the transitions to new languages and techniques with comparison and comprehensive explanations.

INTRODUCTION

I am presenting this to draw attention to the work of the CAMIS group. A full description of the project can be found in the following white paper: https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Data+Visualisation+%26+Open+Source+Technology/WP077.pdf

METHODS

The following table outlines the statistical methods that are actively being researched. For more details, please visit the CAMIS website: https://psiaims.github.io/CAMIS/

Current Methods Actively Researched	
Summary Statistics	Rounding
	Summary statistics
General Linear Models	Students t-test
	Paired t-test
	ANOVA
	ANCOVA
	MANOVA
	Linear Regression
Generalized Linear Models	Logistic Regression
	Poisson/Negative Binomial Regression
	Categorical Repeated Measures

Current Methods Actively Researched	
	Categorical Multiple Imputation
Non-parametric Analysis	Wilcoxon signed rank
	Mann-Whitney U/Wilcoxon rank sum
	Kolmogorov-Smirnov test
	Kruskall-Wallis test
	Friedman test
	Jonckheere test
Categorical Data Analysis	Binomial test
	McNemar's test
	Chi-Square Association/Fishers exact
	Cochran Mantel Haenszel
	Confidence Intervals for proportions
Linear Mixed Models	MMRM
Generalized Linear Mixed Models	MMRM
Multiple Imputation - Continuous Data MAR	MCMC
	Linear regression
	Predictive Mean Matching
	Propensity Scores
Multiple Imputation - Continuous Data MNAR	Delta Adjustment/Tipping Point
	Reference-Based Imputation/Sequential Methods
	Reference-Based Imputation/Joint Modelling
Correlation	Pearson's/ Spearman's/ Kendall's Rank
Survival Models	Kaplan-Meier Log-rank test and Cox-PH
	Accelerated Failure Time
	Non-proportional hazards methods
Sample size /Power calculations	Single timepoint analysis
	Group-sequential designs
Multivariate methods	Clustering
	Factor analysis

Current Methods Actively Researched	
	PCA
	Canonical correlation
	PLS
Other Methods	Nearest neighbour
	Causal inference
	Machine learning

REQUEST FOR CONTRIBUTIONS

Although this project does have a core team, the endeavor of tracking all these comparisons will fail without community contributions. We welcome a wide verity of contributions from correcting small typos all the way to full write-ups comparing software (languages) for a method.

Please contribute by submitting a pull request to and our team will review it. If you are adding a page please follow one of our templates:

- R template

Instructions for Contributions to the CAMIS repository

1. Set up RStudio to clone the CAMIS github repo – See this guidance for more detail
2. If this is your first contribution, contact christina.e.fillmore@gsk.com and give her your github username, requesting to access the CAMIS repo for contributions
3. Go into RStudio and Create a branch –Within RStudio click the branch button (on the git tab top right). Within the box that comes up ensure you are on the “remote=origin” and “Sync branch with remote” is checked. You can name the branch something to do with the amends you intend to make.
4. Edit and /or add files within the CAMIS directories. If you are adding SAS guidance store under sas folder, R guidance store under r folder, for “SAS vs R” comparison store under comp. Follow the naming convention of the files already stored in those folders.
5. Within R studio - Commit each change or new file added, and push to the repo from within R studio.
6. Go into github and do a pull request to sync your branch back to the origin. See create a pull request for more detail. Note that your change will need a reviewer, so please add *DrLynTaylor* and *statasaurus* as reviewers.
7. Once your change is approved, and merged into the origin, the branch will be deleted and you will need to create a new branch to add further contributions. NOTE: you can make the new branch called the same as the old one if you wish but ensure you select to overwrite the previous one.

REFERENCES

Michael S. Rimler, Joseph Rickert, Min-Hua Jen, Mike Stackhouse. 2022. Understanding differences in statistical methodology implementations across programming languages. BioPharm_fall2022FINAL.pdf (higherlogicdownload.s3.amazonaws.com)

FDA Statistical Software Clarifying Statement: https://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM587506.pdf

CAMIS Website: https://psiaims.github.io/CAMIS/

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Brian Varney
Experis, a Manpower Company
Portage, Michigan
Work Phone: (269) 365-1755
Email: Brian.Varney@experis.com