2023 SICASH
SLIIT INTERNATIONAL CONFERENCE ON ADVANCEMENTS IN SCIENCES & HUMANITIES

Testing For Group Differences in Proteomics Data with Left Censored Data and a Limited Sample Size

* [1]P.A.L.A Anurangi, [2]D. Amaratunga, [3]S.D. Viswakula

[1,3]Department of Statistics, Faculty of Science, University of Colombo, Sri Lanka
[2] Princeton Data Analytics LLC, NJ, USA

Corresponding author - *akshila1994@gmail.com

ARTICLE INFO

Article History:
Received: 10 September 2023
Accepted: 01 November 2023

Keywords:
Left censored data; Nondetects; Limit of detection; Proteomic studies

Citation:

P.A.L.A Anurangi, D. Amaratunga, S.D. Viswakula . (2023). Testing For Group Differences in Proteomics Data with Left Censored Data and a Limited Sample Size . Proceedings of SLIIT International Conference on Advancements in Sciences and Humanities, 1-2 December, Colombo, pages 315-319.

ABSTRACT

This research study aims to assess how a specific treatment influences the levels of three proteins when left-censored observations are present in a limited sample size. The dataset contained paired data gathered from 20 subjects categorized into 4 groups with increasing dosages, collected before and after administrating the treatment. The primary objective of this study is to evaluate whether there is an increase in response with increasing dosage for each of the proteins. To check the adherence of data to standard distribution, Cumulative Distribution Function (CDF) plots were used. To obtain summary statistics, Regression on Order Statistics (ROS), Maximum Likelihood Estimate (MLE) and Kaplan-Meier (KM) methods were utilized. ROS assumed to be the estimate that generally works well for the dataset as KM was unable to estimate the median for highly censored data and MLE produced unrealistic values for mean in some cases. Various matched paired tests were used to assess differences between before treatment and after treatment. The censored sign test, censored sign rank test, Peto Prentice test, and censored paired test all produced consistent conclusions across different alternative hypotheses, confirming higher protein concentrations after treatment. To evaluate

mean differences, censored ANOVA, permutation tests, Peto Peto test, and Kruskal Wallis test were employed. No method demonstrated clear superiority over others. Jonckheere Terpstra test revealed the presence of group trend across increasing dosages. Multiple detection limits did not significantly impact the conclusions drawn from the study, and their consideration did not pose additional burdens. In conclusion, the treatment had a significant effect on protein levels, with dose variations influencing the outcome.

1. INTRODUCTION

In laboratory work, readings below limit of detection are a common appearance. These concentrations are called ´nondetects ´or ´left-censored´ and lie between zero and the detection limit of the measuring instrument. (Shoari & Dubé, 2018). According to Aschermann (2008), with the rising number of biotherapeutics are being developed and marketed, the demand for sophisticated analytical methods to characterize therapeutic proteins has driven dynamic developments in protein analysis and proteomics. When a new drug is being developed, it is most often administered to patients as part of a formulation. It may appear straightforward and seemingly simple to accurately quantify the drug substance, but the process can become complex due to the need for robustness and validation. A dataset with a substantial number of undetected values can pose challenges as it can make calculations of descriptive statistics, group differences, correlation coefficients, and regression equations more complex (Helsel and Helsel, 2012). This complexity can result in bias and hinder the ability to draw accurate conclusions from the data. This study analyzes three proteins with nondetects present in a small sample. The primary objective is to test whether there is an increase in response with an increasing dose for each of the proteins.

2. MATERIALS & METHODS

The dataset contained paired data from 20 subjects categorized into 4 groups with increasing dosages, gathered before (time 0) and after (time 1) administrating the treatment. Group 1 acts as the control group (no dose) and Group 2 dosage > Group 3 dosage > Group 4 dosage, respectively. Due to confidentiality, details about the data source or information about the treatments given to the 20 subjects were not available. It is only known that this is from an actual study where the impact of a given drug is measured. At time 0, the protein concentrations are at their natural levels. Time 1 observations are measured after treatment dosages were administrated. A marked increase in any of these protein values indicates a potentially serious side effect of the treatment. Overall censoring percentage of the dataset for the three proteins were 18%, 35% and 35% respectively. Observations within a group may have multiple detection limits depending on the instrument used to measure the values. To evaluate the adherence of data to any standard distribution, the empirical cumulative distribution function (CDF) was compared with theoretical CDFs along with Bayesian information criterion (BIC). Distribution with the lowest BIC value will better fit the data. To obtain summary statistics, Regression on Order Statistics (ROS), Maximum Likelihood Estimate (MLE) and Kaplan-Meier (KM) methods were utilized. Because of the paired nature of the data points, the censored sign test with Fong correction P value, a censored sign rank test with Pratt modification, Paired Prentice Wilcoxon test, Censored Paired test with Q-Q plots were used to assess differences between before treatment and after treatment. To evaluate mean differences, censored ANOVA, permutation tests, Peto Peto test, and Kruskal Wallis tests were employed. Jonckheere Terpstra test was employed to test for group trend. Three proteins were assumed to be independent.

3. RESULTS & DISCUSSION

When determining the adherence of the data to standard distributions, R failed to generate BIC values for protein 1 and protein 2 at time 1 with original data due to an imbalance in datapoints that affected the distance calculations. To balance the impact, numbers were scaled with 1/10 for the above-mentioned instances. Obtained BIC values indicated Gamma distribution tends to go well with data.

At time 0, all three estimates KM, ROS and MLE returned values that are closer to each other for median and mean. At time 1 however, the MLE estimates drastically differ from the rest. A major problem with MLE is that for small data sets, there is often insufficient information to determine whether the assumed distribution is correct or not, and so whether parameters are estimated reliably (Helsel, 2012). Studies have demonstrated that Maximum Likelihood Estimation (MLE) tends to yield poor performance for data sets with less than 25–50 observations. In such cases, alternative estimation methods may be more suitable and accurate (Glelt, 1985). At time 1, KM failed to return estimates for median for protein 1 and protein 2 due to the high censoring percentages that exceeded 50% as the survival step function does not cross the line y=0.5.

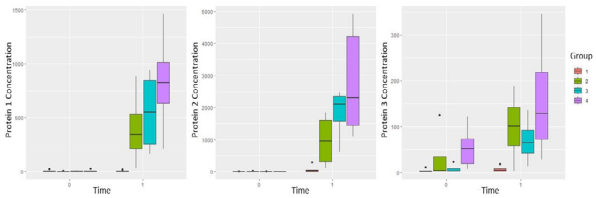

Figure 1 - Grouped boxplots for protein groups & time effect

In this study, the same subject is measured twice before and after treatment which makes it a matched paired case. In the presence of censored data, standard tests like one sample (or paired) t-test fail to precisely determine the differences between pairs that include one or more censored observations.

Table 1: Summary of the results obtained through paired tests

Protein	Censored Sign test (true median difference > 0)	Censored Sign rank test (true difference > 0)	Paired Prentice Wilcoxon Test (Time 1 protein level > Time 0 protein level)	Censored Paired Test (True mean difference > 0))
Protein 1	0.0010*	0.0002*	0.0003*	0.0000*
Protein 2	0.0001*	0.0000*	0.0000*	0.0000*
Protein 3	0.0138*	0.0004*	0.0006*	0.0007*

Hence, suitable tests were selected and utilized. The tests resulted in a significant difference between time 0 and time 1 for all three proteins at 0.05 significance level. Q-Q plots confirmed that for censored paired test, the normality assumption is not obscured. To perform the Kruskal Wallis test, data was re-censored at the highest detection limit. At time 0, ANOVA test and Kruskal Wallis test indicated a significant mean difference in protein 3. Tukey's contrasts for multiple comparison after ANOVA test revealed this to be between group 1 & 4. Due to the limited number of data points, at time 0, R only returned results of Peto Peto test for protein 1. At time 1, all three proteins showed a significant group difference in test results. Multiple comparisons from censored ANOVA revealed that at time 1, the mean differences are between the control group and rest of the dosage groups (2-1, 3-1, 4-1) for the three proteins. Multiple comparisons after Peto Peto test suggested the same for protein 1 and 2, but for protein 3 only 3-1, 4-1 combinations were significant.

Table 2: Summary of the results obtained for group differences.

Protein	Time	ANOVA (log trans-formed units)	Permuta-tion test	Peto Peto	Kruskal Wallis
Protein 1	0	0.965	0.9665 to 0.9648	0.982	0.9824
Protein 2	0	0.574	0.5728 to 0.4848	-	0.7538
Protein 3	0	0.0088*	0.0903 to 0.0972	-	0.0253*
Protein 1	1	0.0000*	0.0051* to 0.0051*	0.0028*	0.0042*
Protein 2	1	0.0000*	0.0067* to 0.0067*	0.0014*	0.0042*
Protein 3	1	0.0003*	0.0249* to 0.0264*	0.0112*	0.0042*

Table 3: Results from Jonckheere Terpstra Trend for Time 0, Time 1 and paired difference

Protein	**P Value Time 0**	**P Value Time 1**	**P Value Paired Difference (Time 0 and 1)**
Protein 1	0.4731	0.0000*	0.0001*
Protein 2	0.3426	0.0000*	0.0000*
Protein 3	0.0042*	0.0001*	0.0061*

The Jonckheere Terpstra trend test can be utilized to ordinal dependent variables instead of the Kurskal-Wallis test when the expected order to the group medians are predetermined. Since R does not provide a specific method to run this test for censored data, few adjustments were made to the data points. All the censored data points were recensored at individual detection limit and then been substituted by the value of detection limit divided by two. Then the paired difference was obtained for each subject between time 1 and time 0. These values were tested at 0.05 significance level. All 3 proteins showed a significant concentration increase moving across groups at 0.05 significance level. Compared to other tests Jonckheere Terpstra test is sensitive to the slightest increases and reductions. It essentially confirms a potential group trend after receiving treatment at time 1.

4. CONCLUSIONS

It can be concluded that the treatment leaves a significant effect on subjects as it increased the protein levels significantly compared to the levels before receiving the treatment. After receiving the treatment, the dosage made a significant difference among the groups. Multiple comparison tests revealed that significant differences were between the control group and the treatment groups but not among treatment dosage groups. When estimating summary statistics for the proteins, ROS worked well in general compared to MLE and KM methods. The matched paired tests used to evaluate the differences between treatment and no treatment (time 0 and time 1) were censored sign test, censored sign rank test, Peto Prentice test and censored paired test produced the same conclusion over different alternative hypothesis confirming that the protein concentrations are higher after treatments are given. To evaluate mean differences, censored ANOVA, permutation tests, Peto Peto test, and Kruskal Wallis test were employed. To confirm existence of a group trend, the Jonckheere Terpstra Trend Test is utilized and the results showed all three proteins having a group trend.

REFERENCES

Antweiler, R. C. (2015). Evaluation of Statistical Treatments of Left-Censored Environmental Data Using Coincident Uncensored Data Sets. II. Group Comparisons. *Environmental Science and Technology*, *49*(22), 13439–13446. https://doi.org/10.1021/acs.est.5b02385

Antweiler, R. C., & Taylor, H. E. (2008). Evaluation

of statistical treatments of left-censored environmental data using coincident uncensored data sets: I. Summary statistics. *Environmental Science and Technology*, *42*(10), 3732–3738. https://doi.org/10.1021/es071301c

Aschermann, K. and Lutter. P. and Wattenberg. A. (2008). Current Status of Protein Quantification Technologies. *Bioprocess International*, *6*, 44–53. https://bioprocessintl.com/upstream-processing/assays/current-status-of-protein-quantification-technologies-182471/

Glelt, A. (1985). Estimation for Small Normal Data Sets with Detection Limits. In *Environ. Sci. Technol,79.*

Helsel, D. R., & Helsel, D. R. (2012). *Statistics for censored environmental data using Minitab and R*. Wiley.

Shoari, N., & Dubé, J. S. (2018). Toward improved analysis of concentration data: Embracing nondetects. In *Environmental Toxicology and Chemistry*, *37* (3). pp. 643–656. Wiley Blackwell. https://doi.org/10.1002/etc.4046