Flipped-data Survival Analysis for Metabolomics Data with Non-Detects

Eric R. Siegel, M.S.[1]
[1]Department of Biostatistics, University of Arkansas for Medical Sciences,
Little Rock, AR 72205

Abstract

Metabolomics data sets typically have a large number of "non-detects": features that are left-censored at the limit of detection (LOD). To analyze data with non-detects, modern methods such as maximum-likelihood estimation and multiple imputation have been deployed in fields as diverse as AIDS research and environmental monitoring. However, such modern methods have yet to be deployed in the metabolomics research arena, where currently prevailing practices for analyzing data with non-detects instead seem to be (A) exclude the non-detects from the analysis, or (B) assign to the non-detects a value such as 0.5 or 1.0 times the LOD. Tekwe *et al.* [1] applied survival-analysis methods to left-censored proteomics data, but their methods were limited to accelerated failure-time models in which parametric distributions were assumed. Here, I apply to left-censored metabolomics data the non-parametric survival-analysis method of Helsel [2], in which data are 'flipped' by subtracting them from a suitably large number, then analyzed using Kaplan-Meier methods and the log-rank test. I use simulation to compare Helsel's method both to currently prevailing practices and to parametric survival regression.

Key Words: Metabolomics, Non-detects, left-censored, Non-parametric

1. Background

Often, the first task in metabolomics analysis is to compare two groups univariately for differences in each metabolic "feature", and to display the results in a "volcano plot" of *P* values versus fold changes. Individual features are then selected for further consideration if their *P* value and fold change satisfy user-defined criteria, such as "*P*<0.05 and fold change >2.0". However, many metabolomics data sets are full of non-detects: features whose intensities are not reported in some samples because they are below the Limit of Detection (LOD). Although such data are obviously left-censored, the current metabolomics literature rarely notes this fact. Most labs restrict analysis to features having <20% non-detects, but provide few further details on how they treat the non-detects they have. In publications, one well-known lab [3] imputed their non-detects as being equal to half the LOD, while another [4] noted the left-censoring, but excluded non-detects from subsequent analysis. Methods that explicitly accommodate the left-censored nature of non-detects could improve the power and accuracy of univariate feature-finding in metabolomics research.

At least two survival-analysis methods can be adapted to metabolomics data with non-detects. One is survival regression, which has already been applied by Tekwe *et al.* [1] to proteomics data with non-detects. This fits the data with parametric error models that assume the equivalent of equal variance between groups. Another is the method of Helsel [2], in which data are 'flipped' by subtracting them from a suitably large number, then analyzed using Kaplan-Meier methods and the log-rank test. Helsel's flipped-data survival-analysis method makes no parametric assumptions about the data.

The purpose of the research reported in this manuscript was to compare Helsel's method both to parametric survival regression and to the common practice of setting non-detects equal to half the LOD. To measure performance, I assessed power to detect a 2-fold change, and I assessed bias when estimating the amount of change.

2. Methods

2.1 Data Simulation

The data-simulation model had the following form:

$$\log_2(W_{ij}) = \mu + 0.5{*}k_i{*}\Delta + \varepsilon_{ij},$$

where W_{ij} represented the true but unobserved (uncensored) data, μ ranged from 12 to 16 in 0.5-unit steps, Δ was either 0 or 1, k_i was –1 and +1 for groups One and Two, respectively, and $\varepsilon_{ij}\sim\text{Nor}(0,2^2)$. At each value of μ and Δ, 50 W_{ij} per group were generated to create a simulated data set, and 1000 simulated data sets were created. The LOD (limit of detection) was defined to be 5,000. The LOD was used to generate the observed data values Y_{ij} and left-censoring indicators I_{ij} from the W_{ij} according to the following rule:

if $W_{ij}\geq\text{LOD}$, then $Y_{ij}=W_{ij}$ and $I_{ij}=0$ (the value was observed at its true value);
if $W_{ij}<\text{LOD}$, then $Y_{ij}=\text{LOD}$ and $I_{ij}=1$ (the value was left-censored at the LOD).

All Y_{ij} were transformed to $\log_2$ units for analysis. The $\log_2$-transformed LOD was 12.2877 for purposes of comparing with the different values of $\mu + 0.5{*}k_i{*}\Delta$ from the data-simulation model.

2.2 Data Analysis

Data analysis was conducted using SAS v9.3 (The SAS Institute, Cary, NC). For t-tests and Kruskal-Wallis tests (and associated effect-size estimates), left-censored values were imputed to be equal to $\log_2(0.5{*}\text{LOD})=11.2877$. The SAS Ttest Procedure was used to estimate the difference in group means and assess it for significance via 2-sided t-test assuming equal group variances. The SAS Npar1way Procedure was used to compute the Hodges-Lehmann estimate of median difference between groups and assess it for significance via 2-sided Kruskal-Wallis test. For Helsel's method, data were "flipped" by subtracting all $\log_2$-transformed data points from the value 30, thereby converting all left-censored data points into right-censored data points. The SAS Lifetest Procedure was used on flipped data to estimate the Kaplan-Meier median of each group, find the difference in the medians, and compare the groups via 2-sided log-rank test for differences in their Kaplan-Meier functions. (If a group experienced ≥50% non-detects, then the median of its flipped data was right-censored at $30.0–\log_2(\text{LOD})=17.7123$, and this value was used to compute the difference of group medians.) For parametric survival regression, data were re-expressed using [lower,upper] syntax [5] to allow direct analysis of left-censored data. The SAS Lifereg Procedure, with "distribution=normal" specified in the Model Statement, was used to estimate the regression parameter for the group difference and assess it for significance using the Wald chi-square test. Under these conditions, the survival regression assumed that data for each group followed normal distributions related purely by location shift, which is equivalent to the equal-variance assumption under which the t-tests were conducted.

To estimate the Type I error of each test procedure when there was no difference between the two groups' means, I tabulated the proportion of simulations in which the test yielded *P*<0.05 when Δ=0 and the number of samples per group was 50. To estimate the power of each test procedure when the two groups' means differed by one $\log_2$ unit, I tabulated the proportion of simulations in which the test yielded *P*<0.05 when Δ=1 and the number of samples per group was 50. Both proportions were accompanied by simulation-based 95% confidence intervals calculated using asymptotic normality.

The bias in each simulated data set was calculated as the estimated group difference minus Δ, both when Δ=0 (no difference between group means) and when Δ=1 (one $\log_2$ unit of difference between group means). The average and standard error of the bias was calculated as the mean and standard deviation (SD) of bias across the 1000 simulated data sets within each combination of μ and Δ. To explore the effect of sample size on bias in estimating Δ, analyses were conducted on each simulated data set using the first 10, the first 25, or all 50 data points per group.

3. Results and Discussion

3.1 Type I Error

Figure 1 shows how the simulated Type I error rates of the test procedures varied with the average non-detect rate in the data; the vertical error bars (sim95%CIs) represent 95%

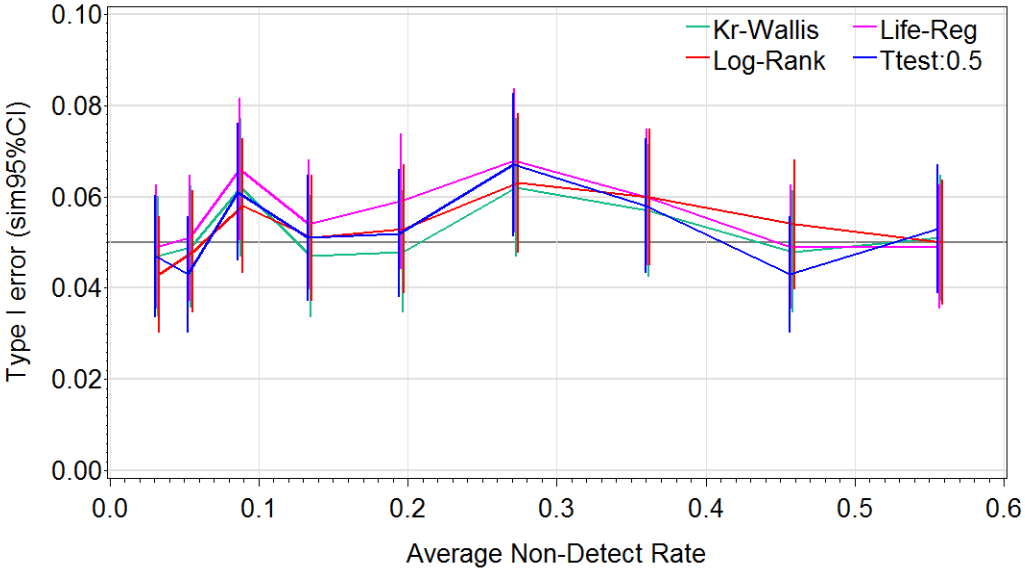

Figure 1. Type I error versus average non-detect rate for the four 2-sided test procedures indicated in the legend when nominal α was 5%, N=50/group, and the true group difference was Δ=0. Vertical error bars represent simulation-based 95% confidence intervals (sim95%CIs) on the Type I error. "Kr-Wallis" = Kruskal-Wallis test on data having non-detects set equal to $\log_2$(0.5*LOD); "Life-Reg" = parametric survival regression with non-detects left-censored at $\log_2$(LOD); "Log-rank" = log-rank test on flipped data with non-detects right-censored at 30–$\log_2$(LOD); "Ttest:0.5" = 2-sample equal-variance t-test on data having non-detects set equal to $\log_2$(0.5*LOD).

confidence intervals on the Type I errors. Although the simulated Type I error rates tended to be above the nominal alpha=5% rate, this nominal rate was excluded by only two of the 36 sim95%CIs shown in **Figure 1**, indicating that, under the conditions examined, the four test procedures had Type I error rates not significantly higher than the nominal alpha=5% rate. Just as importantly, **Figure 1** shows that the four test procedures had Type I error rates that were not significantly different from each other. Interestingly, the four curves of Type I error rate tend to zig-zag up and down together instead of independently. This correlated behaviour suggests that, when Type I error is high, it is high because of a feature of the simulated data, not because of the test procedure.

3.2 Power

Figure 2 shows the simulated power of the four test procedures to declare the group difference significant at 5% alpha when the true difference between groups was Δ=1 $\log_2$ unit, and how this power varies with the rate of non-detects. The vertical error bars (sim95%CIs) represent 95% confidence intervals on the power. Survival regression (magenta curve) uniformly had the highest power at all non-detect rates examined; this result was expected for a survival-regression model with a correctly specified error distribution. Unexpectedly, the log-rank test on flipped data (red curve) had noticeably less power than other methods at non-detect rates of 30% or less. Moreover, in the same range of non-detect rates, the sim95%CIs of the log-rank test show no overlap with those of survival regression, and only a little overlap with those of the Kruskal-Wallis test (green curve). This indicates that the power of the log-rank test on flipped data was significantly lower than that of parametric survival regression. It also indicates that the power of the log-rank test tended to be significantly lower than that of the Kruskal-Wallis test on data that had non-detects set equal to $\log_2$(0.5*LOD), one of several values that made all non-detects tied for lowest rank. Interestingly, the t-test on data having non-detects set equal to $\log_2$(0.5*LOD) (blue curve) had power that was intermediate between survival regression and the Kruskal-Wallis test at all non-detect rates examined.

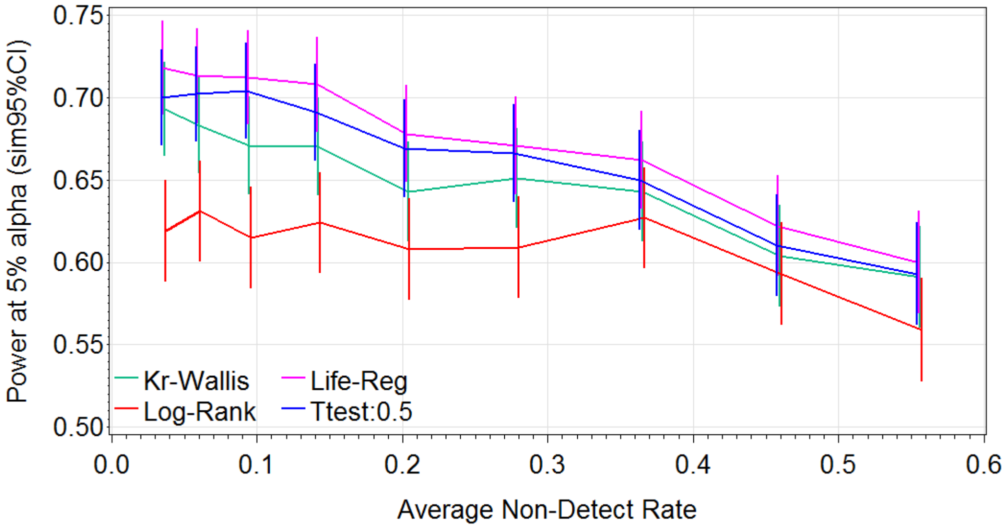

Figure 2. Power versus average non-detect rate for the four 2-sided test procedures indicated in the legend when nominal α was 5%, N=50/group, and the true group difference was Δ=1 $\log_2$ unit. Vertical error bars represent simulation-based 95% confidence intervals ("sim95%CIs") on the power. "Kr-Wallis" = Kruskal-Wallis test on data having non-detects set equal to $\log_2$(0.5*LOD); "Life-Reg" = parametric survival regression with non-detects left-censored at $\log_2$(LOD); "Log-rank" = log-rank test on flipped data with non-detects right-censored at 30–$\log_2$(LOD); "Ttest:0.5" = 2-sample equal-variance t-test on data having non-detects set equal to $\log_2$(0.5*LOD).

Table 1: Bias[1] in Estimating Group Differences When True Difference = Zero

delta=0 log$_2$ units		Difference in group means		H-L[2] Median of Differences		Difference in K-M[3] Medians		Regression[4] Estimate	
% non-detects	**N / group**	**Mean Bias**	**SD Bias**	**Mean Bias**	**SD Bias**	**Mean Bias**	**SD Bias**	**Mean Bias**	**SD Bias**
3.2%	10	-0.02	0.90	-0.02	0.92	-0.04	1.05	-0.02	0.89
	25	-0.02	0.57	-0.03	0.58	-0.02	0.70	-0.02	0.57
	50	-0.00	0.40	-0.00	0.41	-0.01	0.50	-0.00	0.40
8.8%	10	-0.01	0.88	-0.01	0.92	0.01	1.02	-0.01	0.89
	25	0.00	0.58	-0.00	0.58	-0.00	0.68	0.00	0.58
	50	-0.01	0.41	-0.01	0.41	-0.01	0.49	-0.01	0.41
19.6%	10	-0.02	0.83	-0.02	0.88	-0.02	1.03	-0.02	0.89
	25	-0.01	0.56	-0.01	0.56	-0.01	0.71	-0.01	0.59
	50	-0.01	0.39	-0.01	0.36	-0.01	0.49	-0.01	0.41
36.1%	10	0.03	0.77	0.03	0.78	0.01	0.88	0.03	0.95
	25	0.03	0.51	0.02	0.44	0.01	0.67	0.03	0.61
	50	0.02	0.36	0.01	0.24	0.04	0.51	0.02	0.43
55.7%	10	-0.00	0.68	0.01	0.56	-0.01	0.49	0.02	1.50
	25	-0.01	0.40	-0.00	0.20	-0.01	0.26	-0.01	0.64
	50	-0.01	0.29	0.00	0.05	0.00	0.15	-0.01	0.45

1: Bias is calculated as estimate minus delta. The bias mean and SD (standard deviation) was calculated from 1000 simulations per table row. 2: Hodges-Lehmann estimator. 3: Kaplan-Meier estimator (see Methods for when a group had ≥50% non-detects). 4: Survival regression.

3.3 Bias

Table 1 shows the bias mean and SD of the four estimation methods in estimating the difference between two groups when the true difference was zero; note that the SD of the bias is also the SD of the estimator. **Table 2** similarly shows the bias mean and SD of the four estimation methods in estimating the group difference when the true difference was one log$_2$ unit. When there was no difference between the groups, the bias had a mean that

Table 2: Bias[1] in Estimating Group Differences When True Difference = 1 log$_2$ unit

delta=1 log$_2$ unit		Difference in group means		H-L[2] Median of Differences		Difference in K-M[3] Medians		Regression[4] Estimate	
% non-detects	**N / group**	**Mean Bias**	**SD Bias**	**Mean Bias**	**SD Bias**	**Mean Bias**	**SD Bias**	**Mean Bias**	**SD Bias**
3.6%	10	-0.08	0.90	-0.08	0.93	-0.11	1.07	-0.08	0.90
	25	-0.01	0.54	-0.01	0.56	0.00	0.69	-0.01	0.54
	50	0.01	0.39	0.01	0.40	0.00	0.49	0.01	0.39
9.4%	10	-0.06	0.87	-0.04	0.93	-0.03	1.04	-0.04	0.88
	25	-0.04	0.57	-0.02	0.60	-0.03	0.70	-0.03	0.58
	50	-0.01	0.41	0.00	0.43	-0.00	0.50	-0.00	0.41
20.3%	10	-0.03	0.87	0.00	0.93	0.02	1.03	0.04	0.93
	25	-0.09	0.53	-0.06	0.60	-0.03	0.70	-0.03	0.58
	50	-0.08	0.37	-0.05	0.43	-0.02	0.49	-0.02	0.40
36.5%	10	-0.15	0.80	-0.18	0.89	-0.20	0.86	0.06	1.23
	25	-0.19	0.50	-0.26	0.62	-0.13	0.61	-0.01	0.62
	50	-0.16	0.34	-0.25	0.49	-0.04	0.43	0.02	0.42
55.5%	10	-0.36	0.66	-0.54	0.70	-0.63	0.53	0.19	2.02
	25	-0.37	0.40	-0.74	0.47	-0.69	0.35	0.00	0.66
	50	-0.35	0.29	-0.83	0.36	-0.72	0.27	0.02	0.46

1: Bias is calculated as estimate minus delta. The bias mean and SD (standard deviation) was calculated from 1000 simulations per table row. 2: Hodges-Lehmann estimator. 3: Kaplan-Meier estimator (see Methods for when a group had ≥50% non-detects). 4: Survival regression.

was less than ±5% of its SD for all non-detect rates considered, regardless of the N/group, and regardless of the estimation method (**Table 1**). However, when there was a true difference of one $\log_2$ unit between groups, the mean bias varied strongly with both the non-detect rate and the method (**Table 2**). At non-detect rates of 20.3% or less, all four methods had mean biases that were less than ±10% of their SDs at N/group=10, and these means shrank towards zero as the N/group increased. In contrast, at non-detect rates of 36.5% or more, the difference of group means and the Hodges-Lehmann median of differences had mean biases that were negative (indicating underestimation) and either constant or increasing with the N/group. The difference in Kaplan-Meier medians had a mean bias that shrank with the N/group at a 35.5% non-detect rate, but increased markedly with the N/group at a 55.5% non-detect rate; the reason for this behaviour is discussed below. Parametric survival regression showed little bias at all non-detect rates examined.

3.4 Discussion

When I began this research, I originally investigated setting the non-detects equal to three different imputation values commonly found in the environmental-sciences and occupational-hygiene literature: half the LOD (the value used in this report), 0.7071 times the LOD (the LOD divided by the square root of 2), and 1.0 times the LOD. The three different values gave similar results when assessing estimation bias using either the difference of group means or the Hodges-Lehmann median of differences, similar results when assessing Type I error and power of the two-sample equal-variance *t*-test, and identical results when assessing Type I error and power of the Kruskal-Wallis test. The identical results with Kruskal-Wallis test were expected, since all non-detects will be tied for the lowest-ranked value no matter which imputation value is used, so long as that value is ≤1.0 times the LOD. Because overall results were very similar regardless of the imputation value used, I chose to simplify the manuscript, and report results for only the most commonly used imputation value, half the LOD.

Astute readers will notice that I applied the Kruskal-Wallis test from the SAS Npar1way Procedure to my two groups, instead of the equivalent Wilcoxon rank-sum test that was also available. I did so in order to avoid potential confusion: the SAS Lifetest Procedure has a weighted variant of the log-rank test that is also called the Wilcoxon test, and I originally applied this Wilcoxon test as well as the log-rank test to the flipped data. Although I found that the log-rank test on the flipped version of the data (with non-detects treated as right-censored) had significantly inferior power compared to the Kruskal-Wallis test on the unflipped version of the data (with non-detects tied for lowest rank), I also found that the Wilcoxon test on the same flipped data version had power that was stochastically equal to that of the Kruskal-Wallis test on the unflipped data version. This implies that Helsel’s method using the Wilcoxon test had superior power compared to Helsel’s original method using the log-rank test, at least when the two were applied to data whose underlying error components were normally distributed.

Finally, Helsel’s method produced differences in Kaplan-Meier medians that had a mean bias that shrank with the N/group at a 35.5% non-detect rate, but increased markedly with the N/group at a 55.5% non-detect rate. The reason for this was a flaw in my estimation procedure. If a group experienced ≥50% non-detects, then the median of its flipped data was right-censored at $30.0–\log_2(\text{LOD})=17.7123$, and this value was used to compute the difference of group medians. Obviously, the difference of group medians will be underestimated if one of the medians was right-censored, and the chance of having a right-censored group median rises with the non-detect rate. One route to correcting this flaw would be to keep track of the differences calculated from at least one right-censored median, and to exclude them from the bias assessment.

4. Conclusions and Future Research

4.1 Conclusions

At <30% non-detects, the log-rank test on flipped data had less power than the other 3 procedures to detect a 2-fold change between groups. As for estimation, bias was appreciable at >36% non-detects for 2 of the methods, but in the metabolomically relevant regime of ≤20% non-detects, all 4 methods showed mean biases that were small relative to their SDs. These facts indicate that the log-rank test on flipped data may be an inferior procedure when applied to metabolomics data with non-detects.

4.2 Future Research

Parametric survival regression was superior to Helsel's method in this study, presumably because it fit a correctly specified model to normally distributed data adhering to the equal-variance assumption. An obvious direction for future research is to compare how the two methods perform on data whose error components come from the gamma, Weibull, or other distributions. A less obvious, but still important direction for future research is to compare how the two methods perform on data that still is normally distributed, but that violates the equal-variance assumption. Finally, we found preliminarily that the power of Helsel's method at detecting a group difference was improved when a weighted version of the log-rank test was used instead of the unweighted version; future research will explore the conditions on, and limits of, that finding.

References

1. Tekwe CD, Carroll RJ, Dabney AR. Application of survival analysis methodology to the quantitative analysis of LC-MS proteomics data. Bioinformatics. 2012 Aug 1;28(15):1998-2003.
2. Helsel DR. Nondetects and Data Analysis: Statistics for censored environmental data. Hoboken, NJ: John Wiley and Sons, Inc; 2005.
3. Bais P, Moon SM, et al. PlantMetabolomics.org: a web portal for plant metabolomics experiments. Plant Physiol. 2010 Apr;152(4):1807-16.
4. Nicholson G, Rantalainen M, et al. A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection. PLoS Genet. 2011 Sep;7(9):e1002270.
5. The LIFEREG Procedure: MODEL Statement. In: SAS/STAT® 9.3 User's Guide. The SAS Institute, Cary, NC. http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_lifereg_sect011.htm