Weighted Log-Rank Tests for 'Flipped-Data' Survival Analysis of Data with Non-Detects

Eric R. Siegel[1], Songthip T. Ounpraseuth[1], Ralph L. Kodell[1]
[1]Department of Biostatistics, University of Arkansas for Medical Sciences,
4301 West Markham Street #781, Little Rock, AR, 72205-7199

Abstract

Non-detects are data whose values are left-censored at a limit of detection (LOD). Data with non-detects arise in fields as diverse as metabolomics, environmental monitoring, and AIDS research. To analyze data with non-detects, methods such as maximum-likelihood estimation and multiple imputation have been deployed, but these methods require fitting a model whose error term follows the Normal or other parametric distribution. A simple, non-parametric alternative was proposed by Helsel (2005), in which data with non-detects are 'flipped' or converted into right-censored forms by subtracting them from a suitably large number, then analyzed via Kaplan-Meier curves and the log-rank test. In a simulation study, we investigated the performance of Helsel's method on normally distributed data subjected to left-censoring at an LOD. We found that Gehan's generalized Wilcoxon test, a weighted version of the log-rank test, had significantly more power to detect group differences than the standard log-rank test. Here, we explore whether Gehan's test continues to be superior to the log-rank test when the left-censored data are generated using alternatives to the normal distribution.

Key Words: Non-detects, Left-censored, Non-parametric, log-rank, Gehan, power

1. Introduction

Non-detects are data whose values are left-censored at a limit of detection (LOD). Data with non-detects arise in various disciplines such as metabolomics, environmental monitoring, and AIDS research. To analyze data with non-detects, sophisticated methods such as maximum-likelihood estimation and multiple imputation have been deployed, but these methods require the investigator to fit the data to a model whose error term follows the Normal, Logistic, Gamma, or other brand-name parametric distribution. If the incorrect distribution is used, the validity of the investigator's analysis can be compromised. A simple, non-parametric alternative was proposed by Helsel (2005)[1], in which data with non-detects are 'flipped' or reflected into right-censored forms by subtracting them from a suitably large number, then analyzed via standard survival-analysis methods using Kaplan-Meier curves and the log-rank test. Because this method is nonparametric, there is no need to pick a brand-name distribution, and thus no model misspecification. For an exploratory comparison of groups for an endpoint difference in the presence of non-detects, Helsel's method has much intuitive appeal.

But in a recent simulation study[2], we compared several methods, including Helsel's, for their power to detect a difference of 0.5 standard deviations (SDs) between the means of two groups that had normally distributed data subjected to left-censoring at an LOD. We

found that the log-rank test had significantly less power to detect the difference than Gehan's generalized Wilcoxon test, a weighted version of the log-rank test that gives more weight to earlier survival times.

An important feature of our recent study[2] was that the two groups being simulated had normal distributions with equal variances, but different means. In other words, the two groups followed a location-shift model, not a proportional-hazards model. The difference between the two types of model is illustrated in **Figure 1** using "Weibull plots" or plots of ln(–ln(Survival)) versus X, where "Survival" is the survival distribution function of X for each group, and "ln(·)" denotes the natural logarithm of the term inside the parentheses. In a Weibull plot, data that follow a proportional-hazards model maintain constant vertical distance between the two groups' curves, whereas data that follow a location-shift model maintain constant horizontal distance between their curves. Only distributions that form straight lines on Weibull plots can follow both models simultaneously; the only distributions that do this are exponential and Weibull distributions when X is a log-transformed variable, and smallest extreme-value ("Gompertz") distributions when X is the untransformed variable of interest.

In short, by adhering to a location-shift model, the normally distributed data of our recent simulation study[2] automatically violated the proportional-hazards assumption. Allison[3] states that the log-rank test has more power than Gehan's test when the group differences follow a proportional-hazards model, but that Gehan's test "is more powerful than the log-rank test in situations where event times have log-normal distributions with a common variance but with different means in the two groups". This explains why the log-rank test had less power than Gehan's test in our recent study[2], if we assume that Allison meant common variance with different means after logarithmic transformation.

However, this explanation immediately raises two questions. One, would the log-rank test continue to have less power than Gehan's test if the error term in our location-shift model

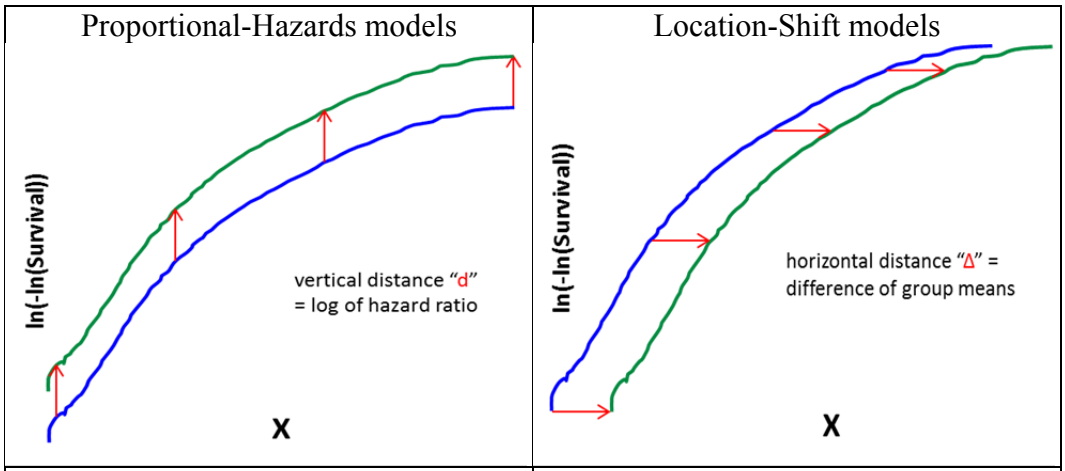

Figure 1: Illustration of of the use of "Weibull plots" to distinguish between data that follow proportional-hazards models and data that follow location-shift (or equal-variance) models. Weibull plots are plots by group of ln(–ln(Survival)) versus X, where "ln(·)" denotes the natural logarithm of the term inside the parentheses. In most Weibull plots, X=ln(Time), but in our Weibull plots, X is more general.

followed parametric alternatives to the normal distribution? Two, how would the log-rank test compare in power to Gehan's test if we simuated data from distributions that simultaneously followed both the location-shift and proportional-hazards model? To answer these two questions, we undertook the simulation study presented in this paper. For our models' error terms, we used the following five distributions: Normal (reference; symmetric), Student's *t* and Logistic (both heavy-tailed symmetric), Gompertz (left-skewed), and Gumbel (right-skewed). For each distribution, we simulated a two-group location-shift model with increasing amounts of difference between the groups. We also simulated increasing rates of left-censoring, in order to continue framing our questions in the context of Helsel's flipped-data method for data with non-detects. We note that the Gompertz and Gumbel distributions have mirror-image density functions, so that flipping Gumbel-distributed data converts it into Gompertz-distributed data, and vice versa.

2. Methods

All simulations and analyses were conducted using SAS v9.3 (The SAS Institute, Inc., Cary, NC). In particular, the LIFETEST Procedure was used to compare flipped data for group differences via the Log-rank and Wilcoxon tests. The SAS documentation for the LIFETEST Procedure states that the Wilcoxon test "is also referred to as the Gehan test or the Breslow test"; hence, we use Gehan's test to mean the Wilcoxon test that is implemented in the LIFETEST Procedure. This is important because Olmsted[4] and others have noted that the LIFETEST Procedure with the Wilcoxon option computes the same rank-statistics vector, but a different covariance matrix, compared to those that would be computed using Gehan's original test procedure.

2.1 Data Simulation

For five distributions d (see below), uncensored data $U_{ij|d}$ were generated as follows:

$$U_{ij|d} = k_i\Delta + \varepsilon_{ij|d},\ \varepsilon_{ij|d} = \zeta_{ij|d} \big/ \sigma_{\zeta|d},\quad i \in \{1,2\},\ j \in \{1,2,...,60\},$$

where i indexed group, j indexed observation within group, Δ (the distance between group means) ranged from 0 to 1 by 0.25, k_i = +0.5 & –0.5 for i = 1 & 2, respectively, and the error term $\varepsilon_{ij|d}$ had SD = 1 by construction. The $\zeta_{ij|d}$ (with SD = $\sigma_{\zeta|d}$) followed the "standard" distribution (with location=0 and scale=1) from one of five location-scale families of distributions d having the following names:

(1) the d=Normal distribution,
(2) the d=Logistic distribution,
(3) the d=Student's t distribution with 8 degrees of freedom,
(4) the d=Gompertz distribution with ζ_{ij}<0 allowed,
(5) the d=Gumbel distribution.

In practice, we generated p_{ij} from the Uniform(0,1) distribution, and computed each of the five $\zeta_{ij|d}$ from the same p_{ij} as:

$$\zeta_{ij|d} = \Phi_d^{-1}\left(p_{ij}\right)$$

where $\Phi_d^{-1}(p_{ij})$ denotes the inverse CDF of p_{ij} for the distribution d in question.

Left-censored data $Y_{ij|d}$ and censoring indicators $I_{ij|d}$ were generated as:

$$Y_{ij|d} = U_{ij|d},\ I_{ij|d} = 0 \text{ if } U_{ij|d} > \text{LOD}_{\text{tCR}|d},\ \text{or}$$
$$Y_{ij|d} = \text{LOD}_{\text{tCR}|d},\ I_{ij|d} = 1 \text{ if } U_{ij|d} \le \text{LOD}_{\text{tCR}|d}$$

where tCR, the target censoring rate, ranged from 0.02 to 0.58 by 0.08, and $\text{LOD}_{\text{tCR}|d}$ was set equal to the tCR^{th} quantile of $\varepsilon_{ij|d}$. For each combination of d, Δ, and tCR, 2500 simulated data sets were generated; the empirical censoring rate was determined for each

such combination as the proportion of censored $Y_{ij|d}$ (i.e., the proportion of non-zero $I_{ij|d}$) aggregated across the 2500 simulations.

2.2 Data Analysis

To implement Helsel's method, both the left-censored $Y_{ij|d}$ and the uncensored $U_{ij|d}$ from each distribution were respectively "flipped" into right-censored forms $X_{ij|d}$ and $W_{ij|d}$ as:

$$X_{ij|d} = C_d - Y_{ij|d}; \quad W_{ij|d} = C_d - U_{ij|d},$$

where the constants C_d were chosen to assure that both $X_{ij|d}$ and $W_{ij|d}$ were always positive. For the overall performance assessment, groups were tested for difference in $X_{ij|d}$ via Gehan's test and the log-rank test at values of Δ ranging from 0.00 to 0.75. To separate more effectively the effect of censoring on power from random fluctuations of the simulation, both $X_{ij|d}$ and $W_{ij|d}$ were tested at Δ=0.50. Power was estimated as the proportion of test results yielding $P<0.05$, and reported with 95% confidence limits that were estimated using asymptotic normality. Adherence of the simulated data to both location-shift and proportional-hazards models was assessed visually at Δ=0.50 for each distribution d by combining all 2500 simulations per d, computing log(–log(Survival)) versus the aggregated $X_{ij|d}$, and plotting the results in Weibull plots.

3. Results and Discussion

(Table 1 and Figures 2–6 have been placed after the References section.)

3.1 Empirical versus Target Left-censoring Rates

Table 1 shows the empirical left-censoring rates obtained for each target left-censoring rate, and how they varied with the error distribution and the distance Δ between group means. As expected, the empirical rate was very nearly equal to the target rate (1) when Δ=0, and (2) when the target rate was 50% on the symmetric error distributions. Otherwise, the empirical rate differed at most by 3.2 percentage points from the target rate, indicating that the simulation's actual left-censoring probability remained close to its target value when Δ was large.

3.2 Operating Characteristics of the Log-rank and Gehan's Tests

3.2.1 Type I Error, all distributions

When Δ=0, power equals Type I error, and power curves become Type I error curves in **Figures 2A** through **6A** (top left panel of each figure). In **Figure 2A**, both tests had stochastically equal Type I error curves, and all but one of the sixteen 95% CIs on the Type I error curves contained the nominal alpha of 0.05. The alert reader will notice that the Type I error curves of **Figures 3A** through **6A** are identical to those of **Figure 2A**. This is because of the way we constructed the $W_{ij|d}$, the censoring thresholds, and the $X_{ij|d}$. When Δ=0, $W_{ij|d} = C_d - \varepsilon_{ij|d}$, the censoring thresholds on each $W_{ij|d}$ lie at $C_d - \mathrm{LOD}_{\mathrm{tCR}|d}$, and $X_{ij|d} = W_{ij|d}$ if $W_{ij|d} < C_d - \mathrm{LOD}_{\mathrm{tCR}|d}$. The different $\varepsilon_{ij|d}$ and $\mathrm{LOD}_{\mathrm{tCR}|d}$ are monotone transformations of the same p_{ij} and tCR, and each such transformation has an inverse. Thus, for fixed i and j, but varying d, the five different $W_{ij|d}$, $X_{ij|d}$, and $C_d - \mathrm{LOD}_{\mathrm{tCR}|d}$ are monotone transformations of each other. This means that, when i and j are fixed while d varies when Δ=0, (1) if the $X_{ij|d}$ are censored for at least one d, then they are censored for all d, and (2) the five uncensored $X_{ij|d}$ have the same rank regardless of d. Because the log-rank and Gehan's tests are rank-based tests, they therefore give the same result on $X_{ij|d}$ for different d when Δ=0, thereby leading to identical Type I error curves for the different d in **Figures 2A** through **6A**.

Of course, these invariance relations are broken when Δ>0, and in consequence, the power curves in **Figures 2A** through **6A** are readily seen to vary with the distribution *d* when Δ>0.

3.2.2 Power, Normal Distribution

Figure 2A plots the simulation power of the log-rank and Gehan's tests versus target left-censoring rate for different values of Δ between two groups whose error terms follow the *d*=Normal distribution. When Δ>0, Gehan's test had higher power than the log-rank test at all censoring rates, and when Δ≥0.50, the 95% CIs on the power curves usually did not overlap for censoring rates <50%. For Δ=0.50, the left panel of **Figure 2B** plots the power of the log-rank and Gehan's tests versus the empirical censoring rate. To separate the effect of censoring rate from random fluctuations of the simulation, the two tests were applied to both the censored $X_{ij|d}$ and their uncensored versions $W_{ij|d}$. Power of both tests decreases as the empirical censoring rate increases, as one would expect. The right panel of **Figure 2B** shows the Weibull plot for the two groups, and confirms that the proportional-hazards assumption fails to hold for this location-shift model with normal-distributed data.

3.2.3 Power, Heavy-tailed Symmetric Distributions

Figure 3A and **Figure 4A** plot the power of the two tests versus target left-censoring rate for different values of Δ between groups whose error terms follow the symmetric heavy-tailed distributions, *d*=Logistic (**3A**) and *d*=Student's t_8 (**4A**). The two figures have very similar-looking power curves. When Δ>0, Gehan's test had higher power than the log-rank test at all censoring rates, and the power curve's 95% CIs did not overlap at low censoring rates. Interestingly, the power of the log-rank test rises towards the power of Gehan's test as the censoring rate increases. The left panels of **Figure 3B** and **Figure 4B** demonstrate that this is not a quirk of the simulation: Power of the log-rank test, when applied to the censored $X_{ij|d}$, increased significantly with the censoring rate relative to its power when applied to their uncensored versions $W_{ij|d}$, at least through 40% censoring. Under the same circumstances, power of Gehan's test began decreasing as the censoring rate increased above 20%. The right panels of **Figure 3B** and **Figure 4B** show the corresponding Weibull plots for these heavy-tailed symmetric distributions. Both plots show curves that are nearly straight and parallel lines at low data values, with pronounced curvature (and pronounced decrease in vertical distance) setting in at high data values corresponding to censoring rates <40%. Thus, most of the violation of the proportional-hazards assumption appears to occur at data values that would not be observed under a 40% censoring rate. At first glance, this suggests that a progressive increase in the censoring rate (up to around 40%) progressively removes more and more of that part of the data that violates the proportional-hazards assumption, thereby leading to an increase in the log-rank test's power. However, it must be noted that the same progressive increase in censoring rate also removes progressively more of those parts of the curves where the vertical distance is relatively small, such that those parts that remain have increased average vertical distance. Since the vertical distance represents ln(HR), the natural log of the hazard ratio between groups, an increase in its average could be the true explanation why the log-rank test's power increased.

3.2.4 Power, Gompertz Distribution before Flipping

Figure 5A plots the power of the two tests versus target left-censoring rate for different values of Δ between groups whose error terms follow the left-skewed *d*=Gompertz distribution before flipping. (After flipping, the error terms become Gumbel-distributed.) For Δ>0, Gehan's test has significantly more power than the log-rank test at almost all

left-censoring rates examined. Strangely, at all censoring rates examined, the power of Gehan's test was approximately constant with censoring rate while the power of the log-rank test rose monotonically with censoring rate. The left panel of **Figure 5B** confirms this behavior at Δ=0.50, and also shows that the power of Gehan's test on the censored version of the data is stochastically equal to its power on the uncensored version of the same data. The Weibull plot of **Figure 5B** explains some of this strange behavior. At low values, the curves are steep and have large vertical distances between them, whereas at high values, the curves are shallow and have small vertical distances between them. As the shallower parts of the curves are progressively trimmed off from the right by more and more censoring, the steeper parts that remain have a progressively larger average vertical distance between them (and thus progressively larger average ln(HR)), thereby increasing power of the log-rank test. Why power of Gehan's test stays constant with censoring rate is harder to explain. However, we note that the shallow regions of the curves correspond to the long right tails of the flipped distributions; these long right tails contribute disproportionately to the group variances. As the increased censoring trims the tails progressively from the right, the variances of what remains will shrink markedly. Perhaps the boost in power that results from shrinking variances is enough to offset the loss of power that one expects from increased censoring, so that the overall power of Gehan's test stays constant with censoring under these circumstances.

3.2.5 Power, Gumbel Distribution before Flipping

Figure 6A plots the the power of the two tests versus target left-censoring rate for different values of Δ between groups whose error terms follow the right-skewed *d*=Gumbel distribution before flipping. Because flipping makes the error terms become Gompertz-distributed, we expect the flipped-data values to adhere to the proportional-hazards assumption. The right panel of **Figure 6B** confirms that they do. In consequence, when $\Delta>0$, the log-rank test has more power than Gehan's test at all censoring rates examined (**Figure 6A**), although the two power curves trend towards convergence at high censoring rates. The left panel of **Figure 6B** shows that power of both tests decreases markedly with censoring rate when each test is conducted on the censored versus uncensored versions of the same data.

4. Conclusions

In our study, the log-rank test was more powerful than Gehan's generalized Wilcoxon test only if the simulated data followed both a proportional-hazards model and a location-shift model simultaneously. If the simulated data followed a location-shift model, but not a proportional-hazards model, then Gehan's test was more powerful than the log-rank test. We note that exponentiation of data that adhere to a location-shift model results in data that adhere to a scale-accelerated failure-time (SAFT) model[5], and that both the log-rank test and Gehan's test are invariant to exponentiation. Therefore, all results on our location-shift models (normal, logistic, student's *t*, Gompertz, and Gumbel) apply equally to their equivalent SAFT models (lognormal, loglogistic, log-t, Weibull, and inverse Weibull, respectively).

Our results also contained a surprise. Power of the log-rank test increased significantly with the censoring rate, at least as first, for the heavy-tailed symmetric distributions, Logistic and Student's *t* with 8 degrees of freedom. And power of the log-rank test actually increased monotonically with the left-censoring rate for the left-skewed (Gompertz) distribution after it was flipped to a right-censored form. Our current thinking is that, in all three cases, the increase in log-rank power with censoring rate is explained

by an increase in the average ln(HR) between groups in the uncensored parts of the distributions.

5. Prospects for Generalization of Results

Can our result of superior Gehan's-test power be generalized to all location-shift (and SAFT) models that violate the proportional-hazards assumption? We believe so, but we are having trouble finding unambiguous support in the literature for this. Collett[6] says simply that the generalized Wilcoxon test is "more appropriate" than the log-rank test for departures from the null hypothesis that do not fit the proportional-hazards assumption, but does not discuss power specifically. The discussion by Allison[3] of which test is more powerful under what assumptions seems to trace back to the seminal paper by Peto and Peto[7]. They considered "Lehmann alternatives" to the null hypothesis, which include proportional-hazards models, and "Normal alternatives" to the null hypothesis, which are location-shift models that follow the normal distribution. To these models, they applied the log-rank test, the probit rank test, and the Wilcoxon rank-sum test. In the absence of censoring, they found the following: (1) asymptotic efficiency of the log-rank test was 100% on Lehmann-alternative models and 82% on Normal-alternative models; (2) asymptotic efficiency of the probit rank test was 82% on Lehmann-alternative models and 100% on Normal-alternative models; and (3) asymptotic efficiency of the Wilcoxon rank-sum test was 75% on Lehmann-alternative models and 95% on Normal-alternative models. Peto and Peto[7] mention that the Wilcoxon rank-sum test is asymptotically efficient on logistic distributions related by location shift, while Kalbfleisch and Prentice[8] state that the Wilcoxon rank-sum test is the optimum rank test if the error distribution is logistic. Kalbfleisch and Prentice[8] go on to state that the generalizations of rank tests to censored data are asymptotically 100% efficient if the assumed model distributions match the "actual" sampling distributions up to location and scaling, which supports our results in the specific cases of the Normal, Logistic, and (flipped) Gumbel distributions. Kalbfleisch and Prentice continue by noting that asymptotic relative efficiency is the square of the limiting correlation of the rank test used with the locally optimum test based on the correctly specified distribution. This, in conjunction with the similar curve shapes in our Weibull plots of **Figures 3B** versus **4B**, suggests that both tests may have nearly the same asymptotic efficiency towards Student's t_8 distribution that they have towards the logistic distribution. However, nothing we have seen in the literature thus far supports a blanket generalization of our results to all location-shift (and SAFT) models that violate the proportional-hazards assumption.

6. Recommendation

In the context of Helsel's flipped-data method for left-censored data, we found that Gehan's test had better power than the log-rank test to detect group differences, for all but one of the distributions we examined. The exception was the Gumbel distribution. In real-world applications, left-censored Gumbel distributions rarely occur. These two facts indicate that, in most real-world applications of Helsel's method, its power to detect group differences would be significantly improved simply by substituting Gehan's generalized Wilcoxon test for the log-rank test.

References

1. Helsel DR. Nondetects and Data Analysis: Statistics for censored environmental data. Hoboken, NJ: John Wiley and Sons, Inc; 2005.
2. Siegel ER. Flipped-Data Survival Analysis for Metabolomics Data with Non-Detects. pp.4156-4162 in: JSM Proceedings, Biometrics Section. Alexandria, VA: American Statistical Association; 2013.
3. Allison PD. Survival analysis using the SAS® System: A practical guide. Cary, NC: SAS Institute, Inc.; 1995.
4. Olmsted A. LIFETEST+ODS+IML=Stratified log rank tests. SAS Conference Proceedings, Western Users of SAS Software (WUSS) 2003. Available at: http://www.lexjansen.com/wuss/2003/DataAnalysis/c-stratified_log_rank_tests.pdf
5. Meeker WQ and Escobar LA. Statistical methods for reliability data. New York, NY: John Wiley & Sons, Inc; 1998.
6. Collett D. Modelling survival data in medical research. Boca Raton, FL: Chapman & Hall/CRC Press LLC; 1994.
7. Peto R and Peto J. Asymptotically efficient rank invariant test procedures. J R Statist Soc A (1972);135(2):185-207.
8. Kalbfleisch JD and Prentice LR. The statistical analysis of failure time data. New York, NY: John Wiley & Sons, Inc; 1980.

Table 1

Table 1: Empirical versus Target rates of left-censoring, as a function of the difference Δ between group means for each of the five error distributions indicated at the far left of the table. Target left-censoring rates are the column headers; empirical left-censoring rates are in the columns under the headers.

	Value of Δ	Target Left-censoring Rate							
		0.02	**0.10**	**0.18**	**0.26**	**0.34**	**0.42**	**0.50**	**0.58**
Normal	**0.00**	0.020	0.101	0.180	0.260	0.340	0.422	0.500	0.582
	0.25	0.021	0.102	0.183	0.261	0.341	0.421	0.501	0.578
	0.50	0.023	0.108	0.190	0.266	0.344	0.422	0.500	0.578
	0.75	0.027	0.115	0.197	0.274	0.349	0.426	0.500	0.574
Logistic	**0.00**	0.020	0.101	0.180	0.260	0.340	0.422	0.500	0.582
	0.25	0.021	0.102	0.184	0.262	0.342	0.422	0.501	0.578
	0.50	0.022	0.108	0.192	0.269	0.347	0.424	0.500	0.576
	0.75	0.025	0.116	0.202	0.280	0.354	0.429	0.500	0.571
Student's t_8	**0.00**	0.020	0.101	0.180	0.260	0.340	0.422	0.500	0.582
	0.25	0.021	0.102	0.184	0.262	0.342	0.421	0.501	0.578
	0.50	0.022	0.108	0.192	0.268	0.346	0.423	0.500	0.577
	0.75	0.025	0.116	0.201	0.279	0.353	0.428	0.500	0.571
Gompertz	**0.00**	0.020	0.101	0.180	0.260	0.340	0.422	0.500	0.582
	0.25	0.021	0.101	0.183	0.261	0.342	0.422	0.503	0.579
	0.50	0.021	0.105	0.189	0.268	0.348	0.427	0.506	0.582
	0.75	0.022	0.109	0.196	0.277	0.357	0.436	0.511	0.584
Gumbel	**0.00**	0.020	0.101	0.180	0.260	0.340	0.422	0.500	0.582
	0.25	0.023	0.104	0.184	0.261	0.341	0.420	0.500	0.577
	0.50	0.032	0.115	0.193	0.266	0.342	0.418	0.495	0.573
	0.75	0.046	0.132	0.205	0.274	0.343	0.415	0.488	0.564

Figure 2

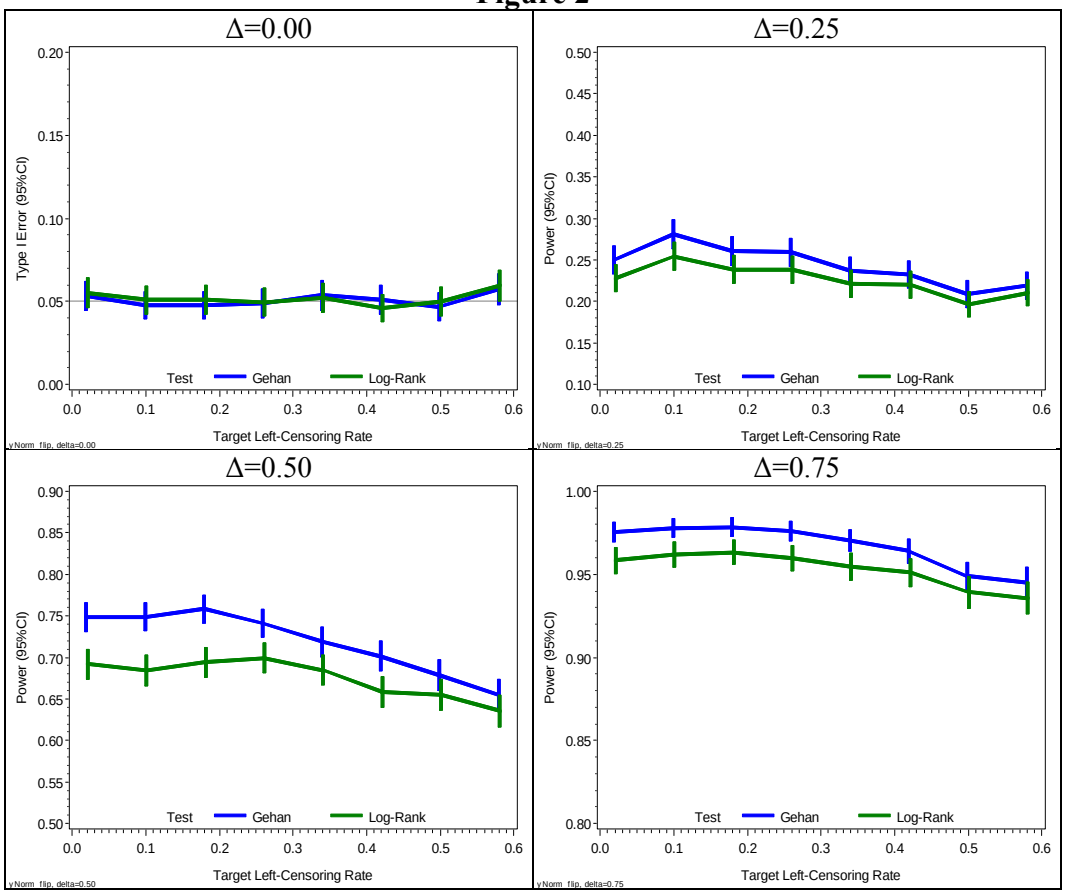

Figure 2A: Power versus test procedure at α=0.05 for left-censored **Normal** distribution flipped by computing $X_{ij}=C_d-Y_{ij}$, where C_d=10. Δ=difference between group means.

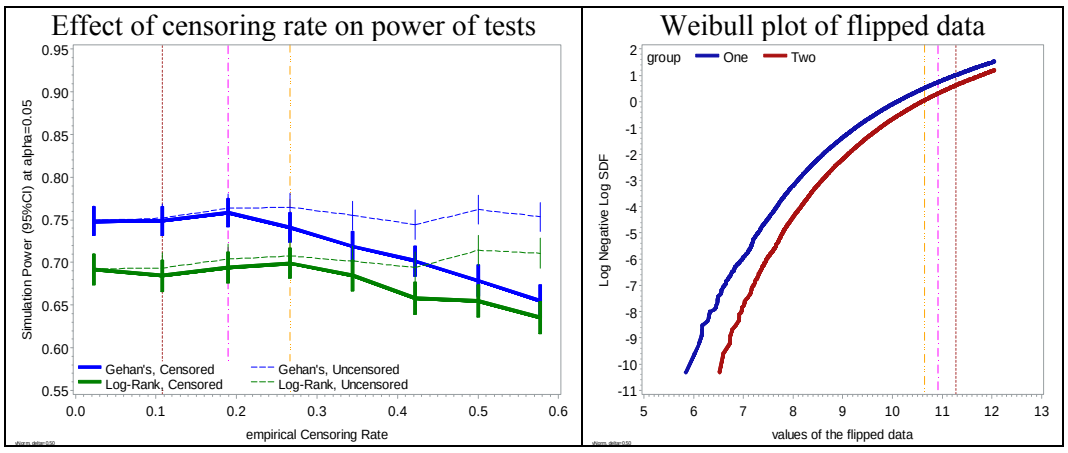

Figure 2B: (left panel) Effect of censoring rate on power of test procedures at Δ=0.50 (compared to uncensored version of same data) when the proportional-hazards assumption is violated. Error distribution is **Normal** before flipping. Δ, the difference between group means, is the constant horizontal distance between curves in (right panel) Weibull plot of flipped data. In both panels, the dashed vertical brown, magenta, and orange reference lines respectively denote empirical censoring rates of 10.8%, 19.0%, and 26.6%.

Figure 3

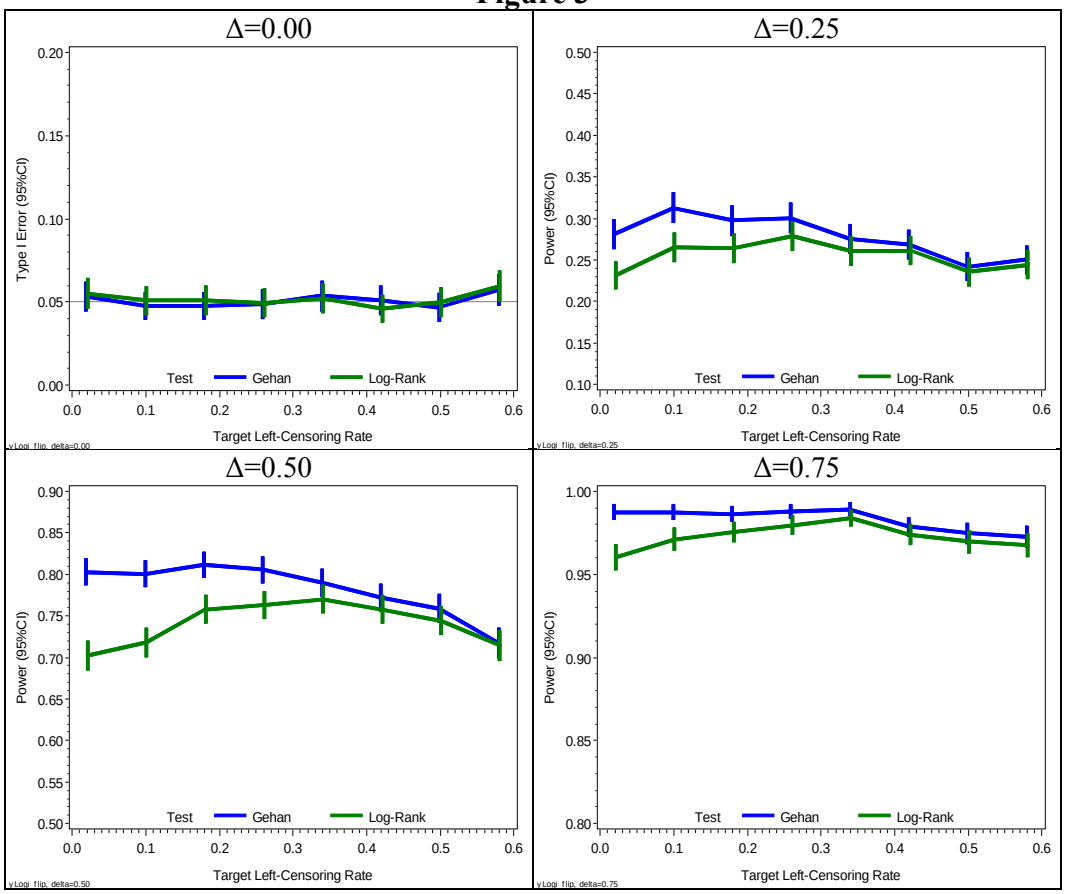

Figure 3A: Power versus test procedure at α=0.05 for left-censored **Logistic** distribution flipped by computing $X_{ij}=C_d-Y_{ij}$, where C_d=15. Δ=difference between group means.

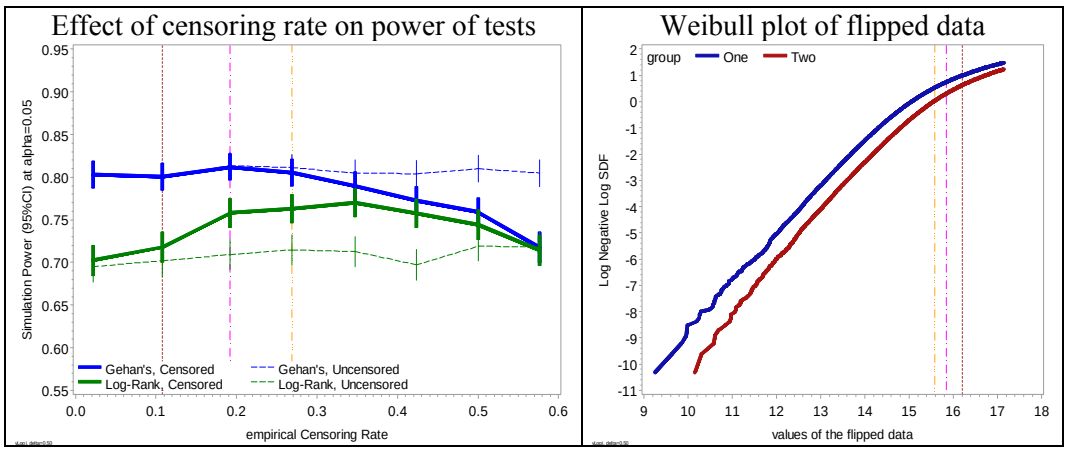

Figure 3B: (left panel) Effect of censoring rate on power of test procedures at Δ=0.50 (compared to uncensored version of same data) when the proportional-hazards assumption is violated. Error distribution is **Logistic** before flipping. Δ, the difference between group means, is the constant horizontal distance between curves in (right panel) Weibull plot of flipped data. In both panels, the dashed vertical brown, magenta, and orange reference lines respectively denote empirical censoring rates of 10.8%, 19.2%, and 26.9%.

Figure 4

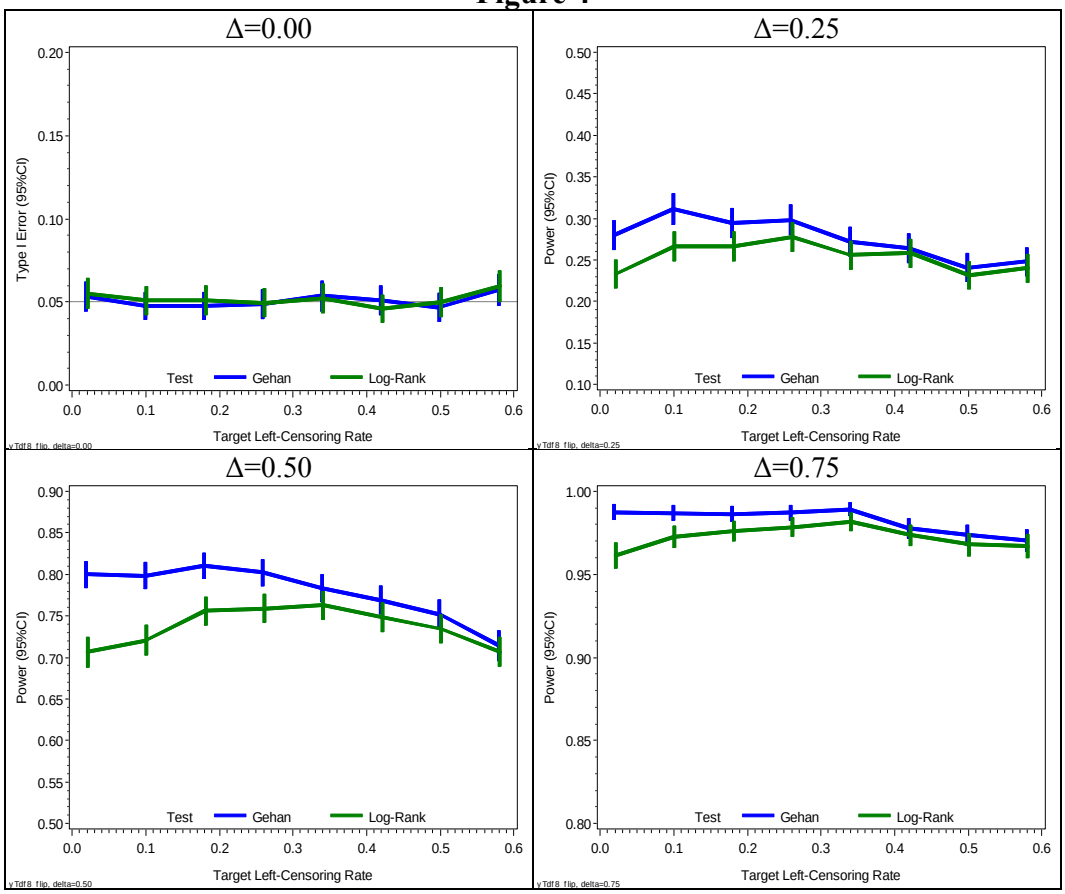

Figure 4A: Power vs test procedure at α=0.05 for left-censored **Student's t_8** distribution flipped by computing $X_{ij}=C_d-Y_{ij}$, where C_d=25. Δ=difference between group means.

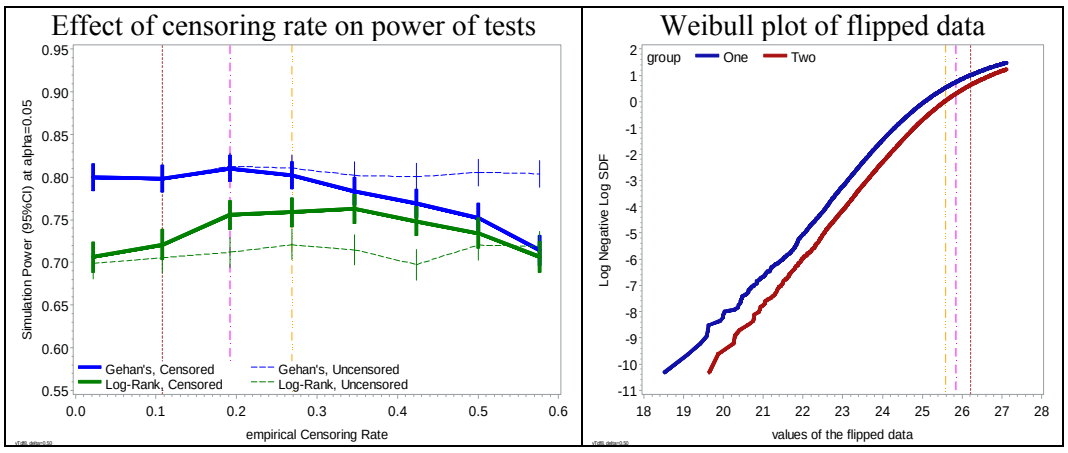

Figure 4B: (left panel) Effect of censoring rate on power of test procedures at Δ=0.50 (compared to uncensored data version) when the proportional-hazards assumption is violated. Error distribution is **Student's t_8** before flipping. Δ, the difference between group means, is the constant horizontal distance between curves in (right panel) Weibull plot of flipped data. In both panels, the dashed vertical brown, magenta, and orange reference lines respectively denote empirical censoring rates of 10.8%, 19.2%, and 26.8%.

Figure 5

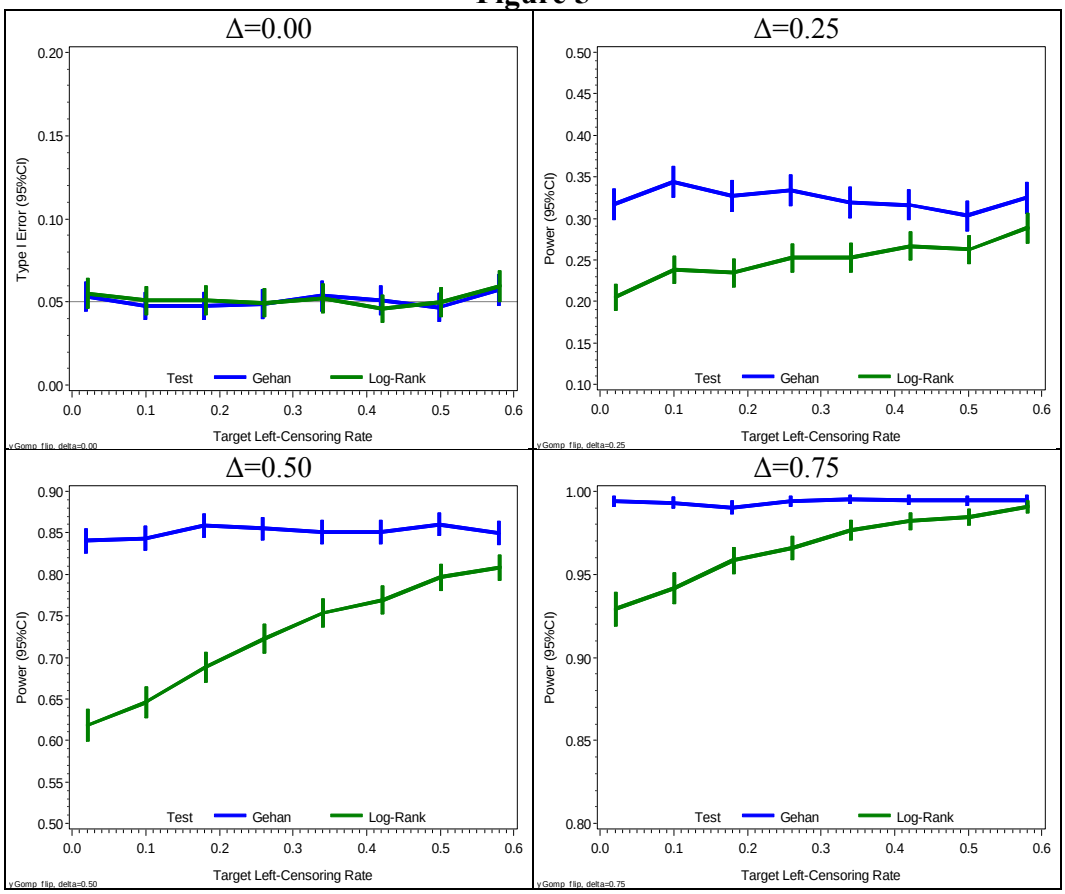

Figure 5A: Power vs test procedure at α=0.05 for left-censored **Gompertz** distribution flipped by computing $X_{ij}=C_d-Y_{ij}$, where $C_d=5$. Δ=difference between group means.

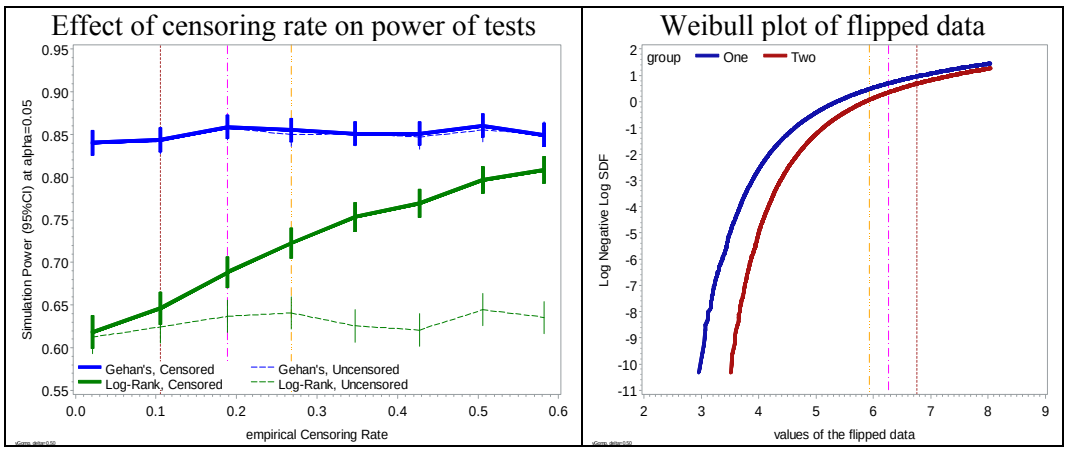

Figure 5B: (left panel) Effect of censoring rate on power of test procedures at Δ=0.50 (compared to uncensored version of same data) when the proportional-hazards assumption is violated. Error distribution is **Gompertz** before flipping. Δ, the difference between group means, is the constant horizontal distance between curves in (right panel) Weibull plot of flipped data. In both panels, the dashed vertical brown, magenta, and orange reference lines respectively denote empirical censoring rates of 10.5%, 18.9%, and 26.8%.

Figure 6

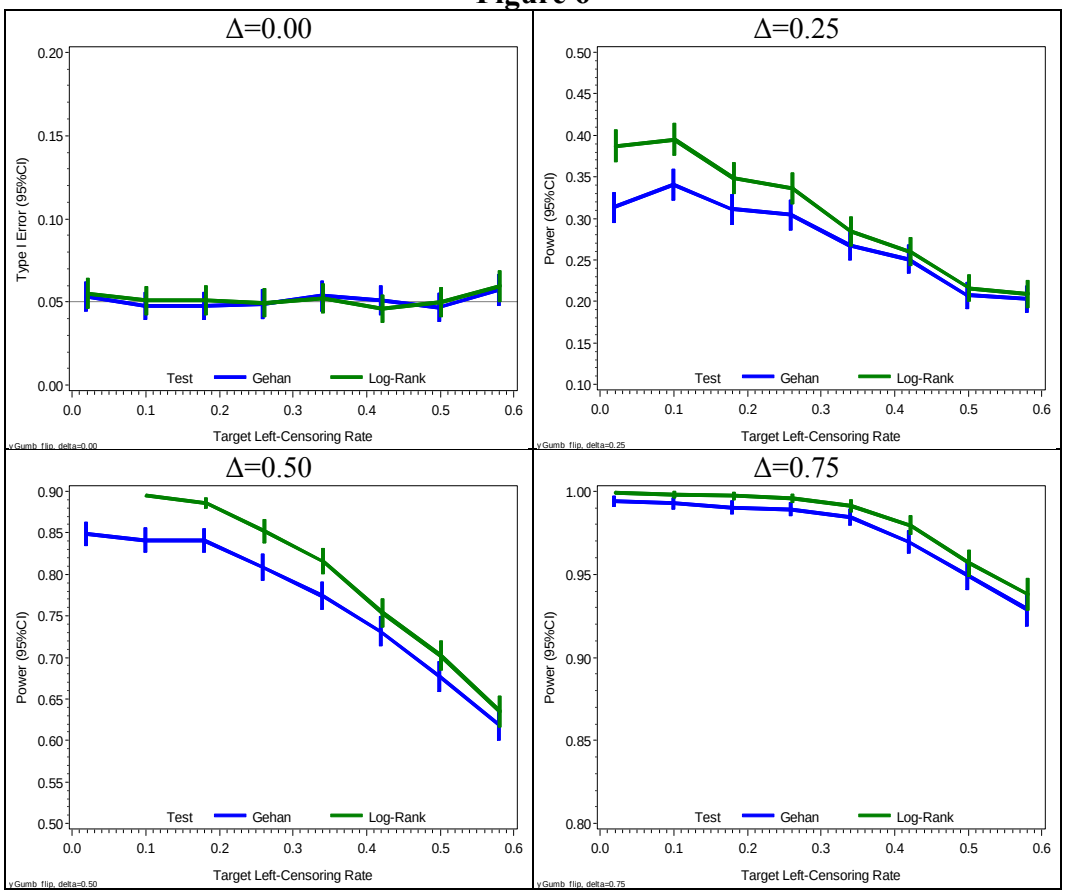

Figure 6A: Power versus test procedure at α=0.05 for left-censored **Gumbel** distribution flipped by computing $X_{ij}=C_d-Y_{ij}$, where C_d=15. Δ=difference between group means.

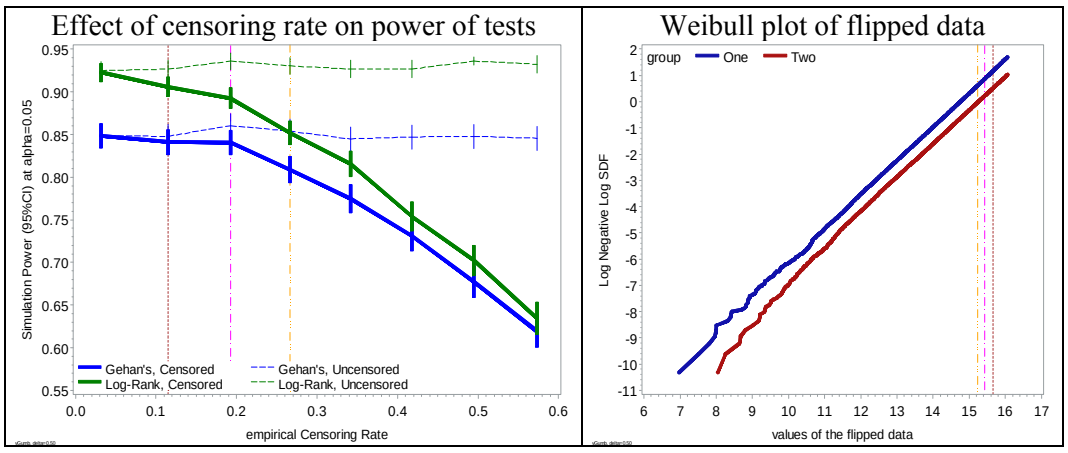

Figure 6B: (left panel) Effect of censoring rate on power of test procedures at Δ=0.50 (compared to uncensored version of same data) when the proportional-hazards assumption is adhered to. Error distribution is **Gumbel** before flipping. Δ, the difference between group means, is the constant horizontal distance between curves in (right panel) Weibull plot of flipped data. In both panels, the dashed vertical brown, magenta, and orange reference lines respectively denote empirical censoring rates of 11.5%, 19.3%, and 26.6%.