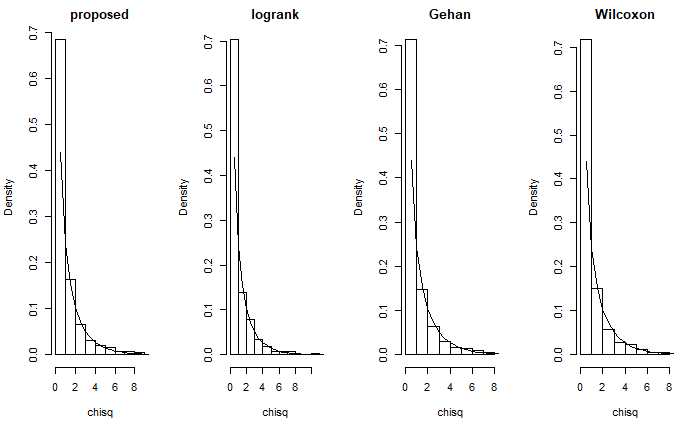
Simulation:

Let the sample size be 200 and the number of replications be 500. The observed data is , where  and . We compare the test statistics of logrank test, Gehan test, Wilcoxon rank-sum test and the proposed test, .

* First case: Let  be independent of , where  and 

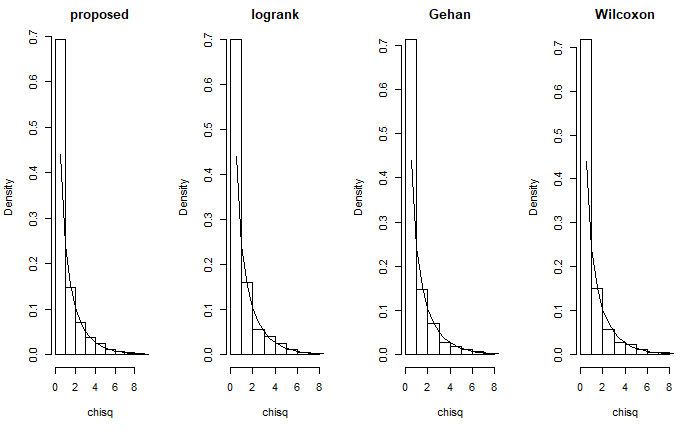
Censoring rate : 0.9



Type I error:

|  |  |  |  |
| --- | --- | --- | --- |
| Proposed | Logrank | Gehan | Wilcoxon |
| 0.06  () | 0.054  () | 0.046  () | 0.048  () |

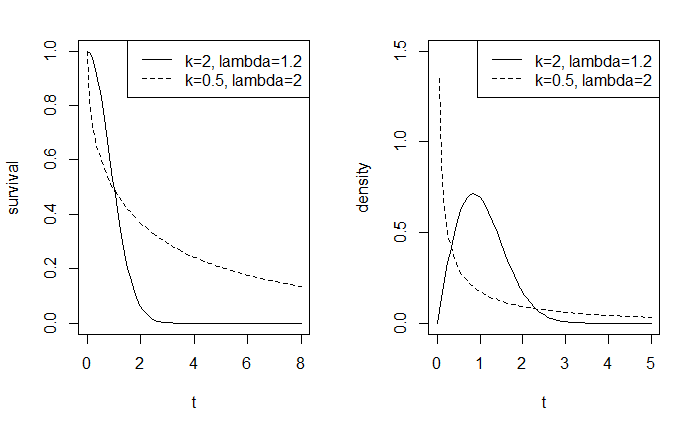
Censoring rate : 0.5



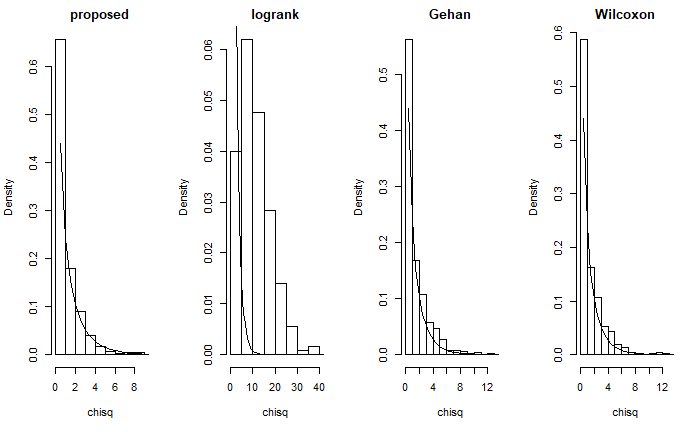
Type I error:

|  |  |  |  |
| --- | --- | --- | --- |
| Proposed | Logrank | Gehan | Wilcoxon |
| 0.052  () | 0.052  () | 0.044  () | 0.048  () |

* Second case: Under this setting, the value of  between  and  is about 0 and the survival functions of these two groups are crossing.



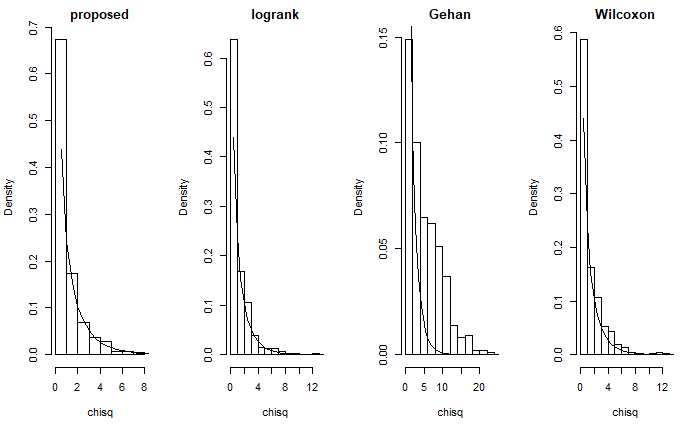
Censoring rate : 0.9



Type I error:

|  |  |  |  |
| --- | --- | --- | --- |
| Proposed | Logrank | Gehan | Wilcoxon |
| 0.04  () | 0.88  () | 0.11  () | 0.096  () |

Censoring rate : 0.5



Type I error:

|  |  |  |  |
| --- | --- | --- | --- |
| Proposed | Logrank | Gehan | Wilcoxon |
| 0.054  () | 0.054  () | 0.516  () | 0.096  () |

* Third case: The value of  between  and  is not 0 and the survival functions are still crossing.

The performance of the logrank test and Gehan test should be poor.

Power: 

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Proposed | Logrank | Gehan | Wilcoxon |
| 0.5 | 1 | 0.668  () | 1 | 1 |
| 0.4 | 0.988  () | 0.186  () | 0.998  () | 0.994  () |
| 0.3 | 0.894  () | 0.062  () | 0.96  () | 0.938  () |
| 0.2 | 0.592  () | 0.23  () | 0.784  () | 0.726  () |
| 0.1 | 0.242  () | 0.614  () | 0.446  () | 0.356  () |

* Fourth case: The value of  between  and  is not 0 and the survival functions are not crossing.

The performance of all of the tests are OK.

Power: 

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Proposed | Logrank | Gehan | Wilcoxon |
| 0.5 | 1 | 1 | 1 | 1 |
| 0.4 | 1 | 1 | 0.998  () | 0.996  () |
| 0.3 | 0.962  () | 0.984  () | 0.946  () | 0.97  () |
| 0.2 | 0.714  () | 0.8  () | 0.7  () | 0.71  () |
| 0.1 | 0.254  () | 0.29  () | 0.254  () | 0.28  () |

* Remarks: Settings

Weibull density: 

First case: Let  be independent of , where  and . For , set . For , set 

Second case: Let  and , . For , set . For , set 

Third case:

1.  For , set .

, 

1.  For , set .

, 

1.  For , set .

, 

1.  For , set .

, 

1.  For , set .

, 

Fourth case:

1.  For , set .

, 

1.  For , set .

, 

1.  For , set .

, 

1.  For , set .

, 

1.  For , set .

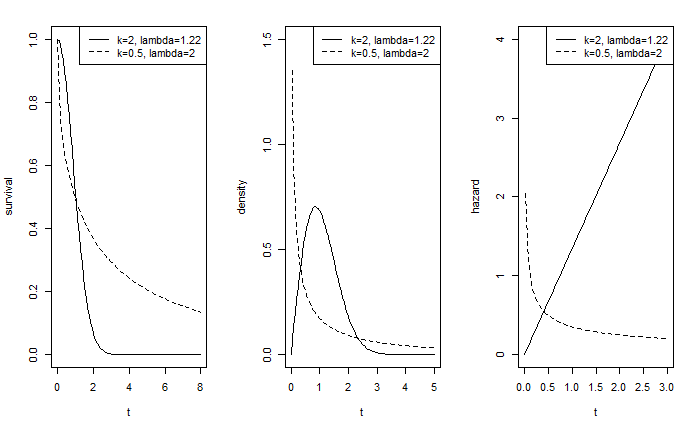
, 

* The difference between two-sample test and testing :

 is not equivalent to .

Example:

Let  and . We can see that these two distributions are quite different but  of  and  is about 0.



The Monte-Carlo simulation is conducted for 500 replications with sample size 200. Then the approximated  is .

Claim:



If we want to test the difference between the distributions, then the power of  would be pretty low when . However, under the null hypothesis, , the type I error of  would still be controlled. I would investigate whether another testing procedure can identify the case that . There are several testing methods to be considered, including log-rank test, Gehan test, Wilcoxon rank-sum test and K-S test.

**Conjectures:**

Under the null hypothesis, , the type I error of all of the testing methods mentioned above could be controlled. For alternative hypothesis, a special case is considered: . Without censoring,  could be written as the form of Mann-Whitney statistic, which is equivalent to Wilcoxon rank-sum test. Therefore, Wilcoxon rank-sum test would not identify the case that . Based on the failure of Wilcoxon rank-sum test, because the first three testing methods take the same approach that assign a score to each observation, these tests might not identify the case mentioned above as well. The last test, K-S test, is effective to test the difference of the distributions no matter what the value of  is. However, the power of K-S test might be poor in another cases.

**Simulation study:**

Under the null hypothesis: 

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Method | Log-rank | Gehan | Wilcoxon | K-S |  |  |
| Type I error | 0.0575  (0.233) | 0.0565  (0.231) | 0.047  (0.001) | 0.043  (0.203) |  | 0.0605  (0.238) |

Under the special case of alternative hypothesis:



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Method | Log-rank | Gehan | Wilcoxon | K-S |  |  |
| Power | 1 | 0.0855  (0.280) | 0.0715  (0.258) | 1 |  | 0.0515  (0.221) |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Method | Log-rank | Gehan | Wilcoxon | K-S |  |  |
| Power | 0.0475  (0.213) | 1 | 0.0715  (0.258) | 1 |  | 0.055  (0.228) |

Under the alternative hypothesis: 

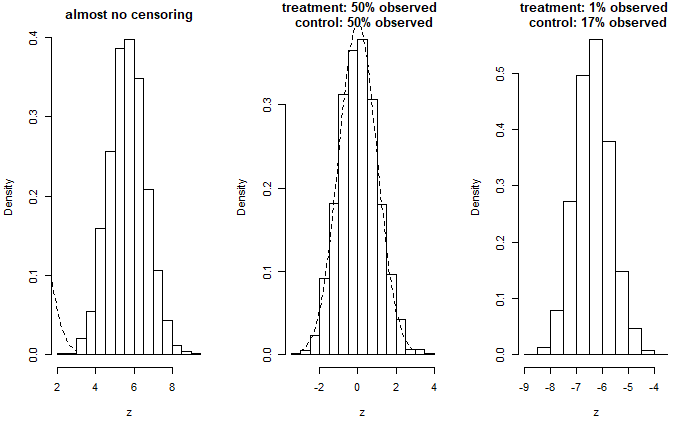
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Method | Log-rank | Gehan | Wilcoxon | K-S |  |  |
| Power | 1 | 1 | 1 |  |  | 1 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Method | Log-rank | Gehan | Wilcoxon | K-S |  |  |
| Power | 0.999  (0.032) | 0.9905  (0.097) | 1 |  |  | 0.9985  (0.039) |

**Comments:**

For Gehan test, the censoring distribution significantly affect the power of testing under the special case of alternative hypothesis. In light censoring case, the performance is almost equivalent to Wilcoxon rank-sum test based on complete data. As the result, the power of Gehan test is pretty low. On the other hand, in heavy censoring case, the difference between the distributions in small  would dominates the testing result. Therefore, it shows that there is a significant difference between the distributions.

The effect of the censoring distribution appears in log-rank test as well. In light censoring case, the difference between the hazard functions can be separated into positive region and negative region, the positive region dominates the log-rank test under our simulation setting. In heavy censoring case, the positive effect and negative effect are cancelled out. Thus, the power of log-rank test is pretty low. If we make the censoring much heavier, then the negative region would dominate the testing result.



**Reviews: Log-rank test, Gehan test**

There are several entry points to discuss the Log-rank test. The first one is assigning a score to each observation and calculate the sum of the scores in treatment group.

For the uncensored observation,  is assigned to the observation,  with . On the other hand, for censored observations,  is assigned to the observation,  with . The conditional expectation assigned to the censored observation has a formula: .

The second one is treating the data as several 2 by 2 tables to calculate the hazards at failure times. The last one is based on the proportional hazard model.

Gehan test is an extended Wilcoxon rank-sum test for censored data.

The validity of logrank test is broken as the survival functions crossing because the hazard functions are crossing as well. (The effectiveness of logrank test depends on the proportional hazard assumption.)

Gehan test: The performance of Gehan’s test might be similar to the test with , but the simulation results show that they are quite different. The difference between these two statistics is the denominator. Gehan’s test is standardized to be normal and the denominator of  is the sample size.

* Survival functions being crossing implies hazard functions being crossing.



