Some thoughts about applying Fleming-Harrington test

Gan Luan Atefeh Javidi

December 11, 2018

1 Abstract

Fleming-Harrington test is a statistical test to test the difference of several survival distributions. This test is used when analysts want to emphasize early, middle, or late departures in survivals. Different emphasis was controlled by two parameters. From the aspect of statistical inference, these two parameters should be determined before performing experiment/clinical trials and observing data. And parameters can be determined based on experience or previous experiment/clinical trial results. However in lacking of experience or previous results, a commonly used but wrong procedure was to investigate data first and then determine parameters and perform the test. We used a data from a lung cancer trial to show this wrong procedure. Simulation was then used to show that real type I error of test used in this procedure can not be well controlled. One way to solve this problem is to split the data into two independent parts with one part used to determine test parameters. After determining the test parameters, the other part of data can be used to do the real analysis. Our simulation results show that type I error is well controlled for test conducted in this manner. However,

power of this procedure is relatively low, since sample size used the formal test is less than the real sample size. Thus another possible way is to invent a new test based on the Fleming-Harrington test and this data splitting method. This new test should control type I error and have a high power. Developing this kind of tests could be the follow up research.

2 Introduction

In clinical studies, the task of comparing the overall equality of two survival distributions with censored observations is a key element in survival analysis. It is well known that the commonly used log-rank test has optimum power under the assumption of proportional hazard rates. However, this assumption is often violated, especially when two survival curves cross each other. A survey demonstrated that under this condition, which was an obvious violation of the assumption of proportional hazard rates, the log-rank test was still used in 70% of studies ([4]). Several statistical methods have been proposed to solve this problem. However, in many applications, it is difficult to specify the types of survival differences and choose an appropriate method prior to analysis.

The presence of delayed effects causes a change in the hazard ratio while the trial is ongoing since at the beginning we do not observe any difference between treatment arms and after some unknown time point, the differences between treatment arms will start to appear. Hence, the proportional hazards assumption no longer holds and both sample size calculation and analysis methods to be used should be reconsidered. The weighted log-rank test allows a weighting for early, middle and late differences through the Fleming-Harrington class of weights, and is proven to be more efficient when the proportional hazards assumption does not hold. The Fleming-Harrington class of weights, allows tuning the two parameters (p, q) depending on if we

expect early, middle or late delays, is proposed in the literature to increase the power at the end of the trial.

The weighted log-rank test is defined as

$$Z_r = \frac{\sum_{i=1}^{D} w(t_i)(d_{1i} - Y_{i1}(\frac{d_i}{Y_i}))}{\sqrt{\sum_{i=1}^{D} w(t_i)^2 \frac{Y_{i1}}{Y_i}(1 - \frac{Y_{i1}}{Y_i})(\frac{Y_i - d_i}{Y_i - 1})}}, \tag{1}$$

Where $z_r \approx N(0,1)$ under the null hypothesis. Fleming and Harrington (1981) proposed the use of w_i to weight early, middle and late differences through the $G^{p,q}$ class of weighted log-rank tests, where the weight function at a time point t_i is equal to

$$w(t_i) = \hat{S}(t_{i-1})^p (1 - \hat{S}(t_{i-1}))^q, \tag{2}$$

Where, $p, q \geq 0$ and $\hat{S}(t_i)$ represents the Kaplan-Meier estimator of survival at time t_i ([2]). Depending on the values of p and q, we will have different weight functions that will emphasize early differences ($p > 0, q = 0$), middle differences ($p = q > 0$) or late differences ($p = 0, q > 0$) in the hazard rates or the survival curves. The parameter combination attributes equal weights to all ($p = 0, q = 0$) data values and hence does not emphasize any survival differences between treatment arms. Moreover, with this parameter combination (1) corresponds to the usual log-rank test.

In classical method since we focus on the entire survival curve rather than the late difference, valid inference requires pre-specification of p and q prior to any data collection. But in modern statistics, after an interim look at the data, the investigator may decide that a different test statistic would be more powerful. Nowadays, this method of making decision about the appropriate test after looking at data is used in a wide range of studies, however, for scientific inference it is not reasonable to look at the survival curves first, then choose weights, as it will increase type I error. We introduce a new method, which split the data to two independent and uncorrelated data set,

use one part to decide about the test and other part to do the test([5]). This way of making decision about the test can control type I error but maybe wasteful in the sense that it is not using all the information in the first subset of the data. Thus a more new test that can use the information from the first part of data is needed.

3 Wrong procedure to analyze survival data

In this section we illustrate one commonly made mistake in analyzing survival data. The data used was taken from "The Statistical Analysis of Failure Time Data" by Kalbfleisch and Prentice, pages 223-224. The data was generated by a clinical trials investigating the survival time of 137 lung cancer patients. This data contains the following 8 variables: treatment type (standard vs test); cell type (squamous, small cell, adeno, or large), survival time (in days), censoring status, Karnofsky score (measure of general performance); patient age (in years), and prior therapy. Patients' age ranges from 34 to 81. These patients are divided into two groups by their age, younger than 60 (group I) and greater or equal to 60 (group II). 53 patients are in group I, while 84 patients are in group II. We want to investigate whether survival are the same for these two groups.

First, survival function of both two age groups were estimated by Kaplan-Meier estimator and plotted below (Fig.1). It is shown that the survival of these two groups were the same for first 80 days. After that the survival of group I is higher than that of group II. Then Fleming-Harrington test with p=0, q=3 was applied to formally compare survival of these two groups since there was a later departure ([1]). p-value of this test is 0.02, thus we concluded that difference exist in survival of these two age groups.

In above procedure, to conduct a formal test to compare survival of two age groups, we first analyzed the data with Kaplan-Meier estimate and no-

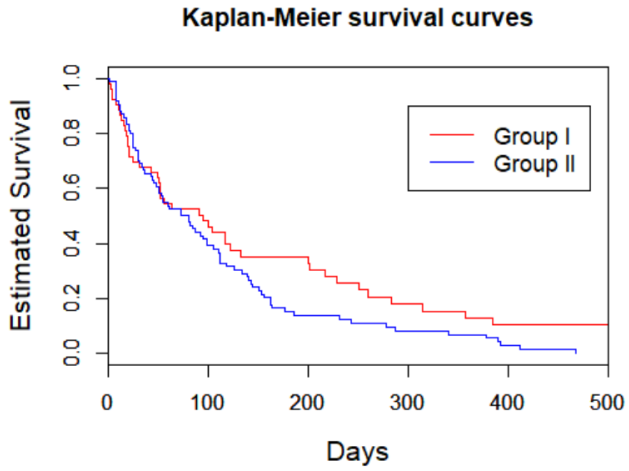

Figure 1: Estimated survival of two survival groups

ticed the existence of late departure.Then we decided to use the Fleming-Harrington (F-H) test focusing on late departure. This kind of similar procedure, choosing test after preliminary analysis of the data, is commonly applied in statistical analysis. However, it is not correct in terms of statistical inference. Next we will show that in general this procedure will increase the type I error by simulation.

4 Simulation result

Simulations were used to show that type I error cannot be controlled if we use the procedure shown above. To simplify the study, we only consider the two sample test. First we investigated type I error in the correct procedure, in which we determine the test parameters before investigating the data. Simulations were run for 5,000 times. For each time, survival times for two groups with same sample sizes of were generated from the same distribution. The censoring indicator was generated from the binomial distribution with a probability of $p = 0.7$. Then tests were applied to test the hypothesis

that survival data of these two groups were from the same survival distribution. The significant level was chosen as 0.05. Then the total number (N) of rejection were counted in all of these 5,000 simulations. Type I error were calculated by N/5000. Four groups of Fleming-Harrington tests were studied. To further simply our study, we only containing the test emphasizing early or late departure and ignoring the test emphasizing middle departure. First group contains two tests with p=1, q=0 or p=0, q=1. Second group contains two tests with p=2, q=0 or p=0, q=2, and the same for the third and fourth group. Log-rank test was used as a control. Two different sample size $n = 40$ and $n = 80$ were investigated. And three distributions were studied: exponential, Weibull, and log-normal distributions. While generating the survival data, other simulation studies in the literature were reviewed ([3]) and most frequently used distributions with their most frequently used parameters were considered for our simulation study. For the exponential distribution, the scale parameter was selected as $\beta = 0.5$; for the Weibull distribution, the shape parameter was $\alpha = 1.5$ and the scale parameter was $\beta = 1$; for the log-normal distribution, the shape parameter was $\sigma = 1$ and the scale parameter was $m = 0$. (Simulation was done in R and Fleming-Harrington test was performed by the function "FHtestrcc" in "FHtest" package.)

Type I error rates for the correct procedure were given in Table 1. From Table 1 we can see that for all three types of distributions, type I errors of Fleming-Harrington tests that emphasizing early departure ($p > 0, q = 0$) are all around 0.05, the desired level. Also no much difference between these two groups of different sample sizes. However, for F-H tests that emphasizing late departure ($p = 0, q > 0$), type I errors are always larger than 0.05. Moreover the departure from 0.05 is bigger when q is larger, with type I be 0.083 for FH(0,4) and n=40. Right now we are clear about the reason for this, and this may deserve further investigation. Not surprising, type I error decrease

for these tests when sample size getting bigger. As a control, Log-rank test has a type I error around 0.05 for in all scenarios. In summary, type I error of F-H tests in the correct procedure are well controlled, though some F-H tests have a slightly higher type I error.

Table 1: Type I error rates for correct procedures

	Exponential(0.5)		Weibull (1, 1.5)		Lognormal(0,1)	
	n=40	n=80	n=40	n=80	n=40	n=80
FH(1,0)	0.0502	0.0554	0.0554	0.049	0.0514	0.05
FH(0,1)	0.0606	0.055	0.0644	0.0538	0.0666	0.054
FH(2,0)	0.0462	0.0442	0.0524	0.047	0.0482	0.0475
FH(0,2)	0.0698	0.0628	0.0724	0.0568	0.0662	0.0645
FH(3,0)	0.0482	0.046	0.05	0.0468	0.0518	0.0505
FH(0,3)	0.077	0.0736	0.0774	0.0634	0.0746	0.0685
FH(4,0)	0.0486	0.482	0.0494	0.047	0.0544	0.0425
FH(0,4)	0.083	0.0732	0.0808	0.0702	0.083	0.074
LR	0.0532	0.0522	0.0538	0.0542	0.056	0.052

FH(1,0) refers to F-H test with p=1, q=0; FH(0,1) refers to F-H test with p=0, q=1 and so on; LR refers to Log-rank test.

Next, we investigated the type I error of F-H test in the wrong procedure. As mentioned above, the wrong procedure means that we determine parameters for test after investigating data not before investigating data. Survival data were generated exactly the same way as explained above. Again we studied four F-H test groups, with each group containing two tests emphasizing early and late departure respectively. First group contains F-H tests with p=1, q=0 and p=0, q=1. Second group contains F-H tests with p=2, q=0 and p=0, q=2, and similar for the rest two groups. To mimic the wrong procedure, for each simulated data, we performed both tests in each group

and select the test that has a lower p value as the final test for that group. Just as above, simulations were run for 5,000 times for each combination of distribution and test group. Type I error was also calculated in the same way as above.

Type I error rates for wrong procedure were given in Table 2. Clearly we can see that type I error in this procedure are much higher than 0.05, which is the desired type I error rate. Also type I error here for each group is higher than that of both tests in the same group in Table 1. Not surprising, type I error decrease as sample size increase from 40 to 80. In summary, this simulation show that type I error cannot be controlled if we determine the parameters after investigating the data.

Table 2: Type I error after investigating the data

	Exponential(0.5)		Weibull (1, 1.5)		Lognormal(0,1)	
	n=40	n=80	n=40	n=80	n=40	n=80
FH(1,0)	0.106	0.0972	0.0994	0.084	0.105	0.0915
FH(0,1)						
FH(2,0)	0.1152	0.1016	0.1098	0.1075	0.1184	0.1105
FH(0,2)						
FH(3,0)	0.1154	0.1098	0.1226	0.1138	0.125	0.116
FH(0,3)						
FH(4,0)	0.128	0.1174	0.1246	0.1202	0.123	0.119
FH(0,4)						

FH(1,0) refers to F-H test with p=1, q=0; FH(0,1) refers to F-H test with p=0, q=1 and so on.

From simulations we have shown that in order to control type I error, we need to determine the p and q for F-H test before investigating data. If analysts have an expectation for the early, middle, or late departure, or

they more focus on a certain kind of departure, they may determine the p, q values accordingly. Also analysts may have scientific reasons to believe a certain kind of departure or similar previous study is available to determine departure. In both cases, p, q can be determined easily. However, there are some situations none of these information are available and analysts still want to use F-H test. Then the question is how should they determine p, q to well control type I error and make power of test as high as possible. One way is to split the data into two independent parts. One part of data is used to determine p and q and then formal test is performed with the rest of data. Next we use simulation to investigate whether type I error is well controlled in this situation.

In this simulation, survival data were generated exactly as shown above and also same four groups of F-H tests were investigated. In each simulation run, survival data were split randomly into two parts. One contains 20% of the data while the other contains 80% of the data. First both tests from the same group were applied to the 20% data. Test that has a smaller p values was selected. Then the formal test were performed with the selected test on the rest 80% part of the data. Just as above, simulations were run for 5,000 times for each combination of distribution and test group. Type I error was also calculated in the same way as above.

Type I error rates of tests from data splitting procedure were shown in Table 3. We can see that all of these type I errors are well controlled, tough some of them are slightly higher than 0.05. For almost all of these groups, type I error decrease as sample size increasing from 40 to 80. We also noticed that type I error increase as p/q values increase. This is easy to explain. This group type I error here can be treated as the some kind of average of the type I errors of two tests in same group in Table 1. And in Table 1, type I errors increase as q increasing for F-H test with $p = 0, q > 0$ and type I errors are all round 0.05 for F-H test with $p > 0, q = 0$. In summary, this

data splitting method can well control type I error. This result is expected since two subparts of the data were independent, thus investigating one part of the data does not affect the test applied on the other part of the data.

Table 3: Type I error after data splitting

	Exponential(0.5)		Weibull (1, 1.5)		Lognormal(0,1)	
	n=40	n=80	n=40	n=80	n=40	n=80
FH(1,0)	0.0626	0.0556	0.054	0.0554	0.0732	0.0565
FH(0,1)						
FH(2,0)	0.0664	0.0618	0.067	0.0536	0.071	0.0805
FH(0,2)						
FH(3,0)	0.0674	0.0632	0.066	0.0635	0.0718	0.076
FH(0,3)						
FH(4,0)	0.075	0.0628	0.0712	0.0608	0.0784	0.0745
FH(0,4)						

5 Discussion

In this project, we show that the commonly used statistical method, determining test parameters after observing data, is not correct. This method cannot control the type I error. Thus the correct way to perform the test is to determine test parameters before observing data. In the case that test parameters can not be determined before observing data, the data splitting methods can be used. We also show that this data splitting method can control type I error by simulation.

One important issue that was not investigated in this project is power. As mentioned above, when we use the data splitting method, usually test power will decrease, since test power is very sensitive to sample size. By changing

the ratio of the two subsets in splitting, we may be able to find the ratio that has the highest power. This requires further investigation. Another way to increase the test power is to reuse the subset of the data that used to determine the test parameters. Of course, we cannot use this subset of data in the same way as for the rest of the data not be used before. A new test that treats these subsets of data differently needs to be promoted. For this new test, we want type I error to be controlled and still have the highest possible power. Developing this new test could also be a further research direction.

Actually this is not the only situation in statistical analysis that analysts looking at data first and then decide about the test. One example is the linear regression with multiple covariates. Analysts usually build the model by selecting some significant covariates and they perform the statistical inference on coefficients of these selected covariates. This is exactly the procedure that determine the test parameters (or model variables) after observing the data. Another example is that uasually several models are available for analyzing the data. Sometimes what people do in this situation is to fit all the models and select the models that fit the data best. Generally, type I error cannot be well controlled for test under these procedures. All these examples are special cases of the research area of selective inference. This is a relatively new area and needs more study.

References

[1] Gars, V., Andrieu, S., Dupuy, J., Savy, N. (2018) *On the FlemingHarrington test for late effects in prevention randomized controlled trials.* arXiv:1806.11294v2

[2] Jimenez,J., Stalbovskaya, V., Jones, B. (1980) *Properties of the weighted*

log-rank test under delayed effects assumption in the design of confirmatory studies with delayed effects. Technometrics 22, 325331, 1980.

[3] Karadeniz, P.G., Ercan, I. (2017) *Examining test for comparing survival curves with right censored data*. Statistics in Transition, Vol. 18, No. 2, pp. 311328, DOI 10. 21307

[4] Kristiansen IS. (2012) *PRM39 Survival curve convergences and crossing: a threat to validity of meta-analysis? Value in health* 15(7): A652 doi: 10.1016/j.jval.2012.08.290

[5] Maya, R.J., Maier b, H.R., Dandy, G.C. (2010) *Data splitting for artificial neural networks using SOM-based stratified sampling*. Neural Networks 23 (2010) 283294.