

Some Thoughts about Applying Fleming-Harrington Test

Gan Luan Atefeh Javidi

Department of Mathematical Sciences, NJIT

December, 11, 2018

- Introduction
- Wrong procedure to analyze survival data
- Simulation study
- Discussion

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. 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, hence, the proportional hazards assumption no longer holds and both sample size calculation and analysis methods to be used should be reconsidered.

Handel problem

Using weighted log-rank test

The weighted log-rank test allows a weighting for early, middle and late differences through the Fleming and Harrington class of weights, and is proven to be more efficient when the proportional hazards assumption does not hold. The Fleming and 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.

Weighted log-rank test

The weighted log-rank test is defined as

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

Where $z_r \approx N(0, 1)$ under the null hypothesis.

Weighted log-rank test

Fleming and Harrington (1981) proposed the use of $W(t_i)$ 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

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

Where, $p, q \geq 0$ and $\hat{S}(t_i)$ represents the Kaplan-Meier estimator of survival at time t_i .

Different weight functions depending on the values of p and q

- Early differences ($p > 0, q = 0$)
- Middle differences ($p = q > 0$)
- Late differences ($p = 0, q > 0$)
- Log-Rank ($p = 0, q = 0$)

Classical method vs modern method

Classical method

- Select hypotheses/model/question
- Collect data
- Perform inference

Modern method

- Collect data
- Select hypotheses/model/question
- Perform inference

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 next illustrate this wrong procedure by analyzing a real clinical trial data.

Clinical Trial data

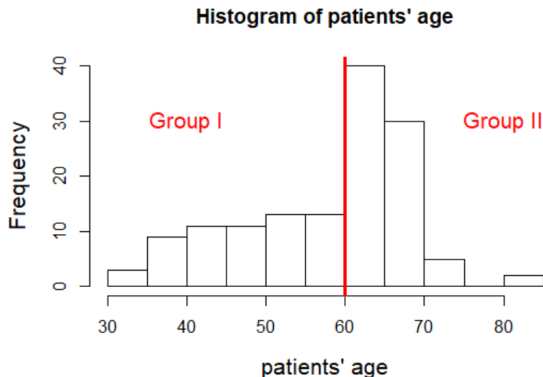
The data used was taken from "The Statistical Analysis of Failure Time Data" by Kalbfleisch and Prentice, pages 223-224. The data was generated by a clinical trials investigating the survival time of 137 lung cancer patients.

Data containing

- treatment type (standard vs test)
- cell type (squamous, small cells, adeno, or large)
- survival time (in days)
- censoring indicator
- Karnofsky score (measure of general performance)
- patient age (in years)
- prior therapy

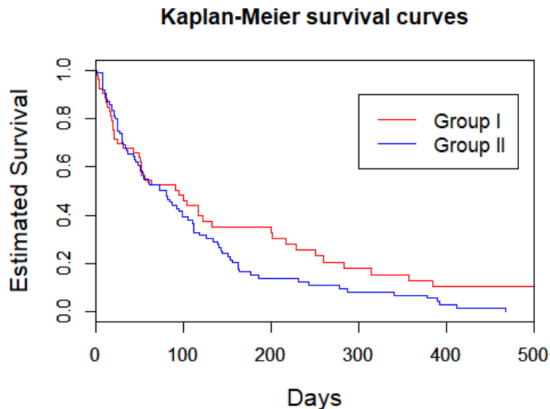
Defining a new variable: age group

Patients' age range from 34 to 81.



We try to investigate whether patients survival were the same for these two age groups.

Kaplan-Meier estimator



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.

Fleming-Harrington Test

Since late departure exist in survival of these two groups, we decided to apply the Fleming-Harrington test that emphasizing late departure. We use $p=0$, $q=3$.

p-value of the test is 0.02, thus we concluded that difference exist in survival of these two age groups.

WRONG PROCEDURE

- First investigating data
- Then choosing statistical test based on data
- Last draw conclusion

Next we use simulation to show that this method will actually increase type I error.

Simulation study

- Simulations were used to show that type I error cannot be controlled if we use the procedure shown above.
- To simplify our study, we only consider two sample test.
- First we investigate the type I error in correct procedure, then investigate that in wrong procedure, and compare them.

Simulation setup- type I error in correct procedure

- Simulations were run for 5,000 times.
- For each run, survival times for two groups with same sample sizes ($n = 40$ or $n = 80$) were generated from the same distribution (Exponential (0.5), Weibull(1, 1.5), lognormal(0,1)).
- 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 of rejections was counted in all of these 5,000 simulations (denoted as N). Type I error were calculated by $N/5000$.

Simulation setup – type I error in correct procedure

To further simplify, we only consider Fleming-Harrington tests that emphasize early or later departure and ignore tests that emphasize middle departure.

- Four groups of Fleming-Harrington tests were studied.
 - $p=1, q=0$ or $p=0, q=1$
 - $p=2, q=0$ or $p=0, q=2$
 - $p=3, q=0$ or $p=0, q=3$
 - $p=4, q=0$ or $p=0, q=4$
- Log-rank test was used as a control.

Simulation result - type I error in correct procedure

	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

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.

Simulation setup - type I error in wrong procedure

- Survival times and censoring indicators were generated same as above in the correct procedure.
- Four groups of Fleming-Harrington tests were studied.
 - $p=1, q=0$ or $p=0, q=1$
 - $p=2, q=0$ or $p=0, q=2$
 - $p=3, q=0$ or $p=0, q=3$
 - $p=4, q=0$ or $p=0, q=4$
- 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.
- simulation was run for 5,000 times and type I error rate was calculated the same as shown above.

Simulation result - type I error in wrong procedure

	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)						

summary

This simulation show that type I error cannot be controlled if we determining the parameters after investigating the data.

Data splitting method

- 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.
- p , q can be determined based on the experimenter's expectation for a certain type of departure, based on some scientific reasons, or similar previous study.
- When there is no information available to determine p and q , the question is how should they determine p , q to well control type I error and make power of test as high as possible.

Data splitting

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.

Simulation setup - Type I error in data splitting

- Survival data were generated exactly same as shown above.
- 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.
- Both tests from the same group were applied to the 20% data. Test that has a smaller p values was selected.
- The formal test was performed with the selected test on the rest 80% part of the data.
- Simulations were run for 5,000 times and Type I error was also calculated in the same way as above.

Simulation result - type I error in 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)						

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.

- One important issue that was not investigated in this project is power.
 - Data splitting method usually reduce the power.
 - Different ratio for data splitting may have different power.
 - New method to reuse the part of data that was used to determine p , q values.
- This wrong procedure talked about in this project is not the only scenario that analysts looking at data first and then decide about the test/model.
 - Linear regression in multiple covariates
 - Data analysis with several models/tests available

Reference

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



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.

Thank You!