Comparing two survival function in the case of heavily censored data

Ahmed Hamimes[*1], **Benamirouche Rachid**[2]

[1] PHD Student, National Higher School of statistics and applied economics, Tipaza, Algeria,ahmedhamimes@yahoo.com.

[2] Professor, National Higher School of statistics and applied economics, Tipaza, Algeria,,rbena2002@hotmail.com

Received: 11/10/2020 *Accepted: 26/12/2020* *Published:31/12/2020*

Summary

To test for equality between heavily censored survival functions. The comparison between the average risk values is used. This value calculated in the Kaplan Meier model, according to a Bayesian design, and through the posterior mean approach. . This method gives credibility to the results found because the calculation of the risk mean is not identical for all durations because of incomplete data or censored data.

Keywords: the posterior mean approach, Kaplan Meier model.

JEL Classification Codes : C11, C12, C41.

Résumé

Pour tester l'égalité entre les fonctions de survie fortement censurées. On utilise la comparaison entre les valeurs moyennes des risques. Cette valeur calculée dans le modèle de Kaplan Meier et selon une conception bayésien et à travers l'approche de la moyenne a posteriori. Cette méthode donne une crédibilité aux résultats trouvés car le calcul de la moyenne de risque n'est pas identique pour toutes les durées à cause des données incomplètes ou bien les données censurées.

Mots clés : l'approche de la moyenne a posteriori, le modèle de Kaplan Meier.

Codes de classification de JEL : C11, C12, C41.

[*] ***Corresponding author, e-mail :*** ahmedhamimes@yahoo.com

1. Introduction

Lifespan research consists of analyzing the occurrence of an event over time, whatever its nature (death, onset of a disease, recurrence of a disease, etc.). For this, it is important to have the time to follow each topic as well as when the incident occurred. The peculiarity of these studies for participants who did not ask the occurrence in question at the time of the research report is [the presence of incomplete data, called censored data; this requires an appropriate technique for their analysis]. Thus, research on the lifespan finds applications in many fields, whether they are biomedical or not, such as industry (for studies of the reliability of systems), physics (with the study of the duration of particle life) ... The first models for analyzing survival times were developed in order to model the observed survival in a univariate manner. They are nonparametric models like those used in the process of Kaplan and Meier (1958). Given the clinical and epidemiological needs to simultaneously take into account several variables, parametric models have been proposed to impose an a priori distribution of survival data (e.g. exponential distribution, Weibull distribution, etc.) in order to explain the effects. prognostic factors in a manner similar to that used in multiple regression (Feigl and Zelen, 1965). The methods of studying survival were initially built on the basis of inferential statistics, that is, according to a frequentist approach. Thanks to the work of Reverend Thomas Bayes (Barnard, 1958) and his essay "An Essay to Solve a Problem in the Doctrine of Chance," a radically new approach to statistics was born. Thomas Bayes exposes the essence of conditional probabilities in this text, but faced with a logic allowing to deduce the probabilities of a given cause, Thomas Bayes looked at the inverse problem of the evaluation of the causes of observable events, supposed to be unknown Therefore, Bayesian methods in statistics have their roots in these ancient works; however, it should be noted that they were not widely used in the field of biomedicine until the early 1980s. Relatively recent advances, both at the theoretical level, with Markov chain theory and the The implementation of efficient and functional sampling methods, with increasingly powerful means of measurement, are at the origin of the increase in the production and use of these methods. It is normal that many methods have been developed for Bayesian survival analysis. Therefore, users of survival analysis methods generally have to select one tool from the range available. This choice does not only depend on the form of survival to be measured, but also on the statistical approach to inference used. In the case of semi-parametric models, many researchers have explored the use of Bayesian inference ((Ferguson, 1973; (Kalbfteisch, 1978;

Florens and Rolin, 2001) and parametric (Feigl and Zelen, 1965); (Carlin, Gelfand and Smith, 1992), etc. Such works have been constructed in different methodological contexts by using specific a priori distributions and / or by modeling the cumulative risk function, or more precisely the instantaneous risk function.

2. Kaplan Meier model

Nonparametric models make it possible to estimate one of the different functions characterizing the distribution of the variable X without making any a priori assumption about it. In this approach, it is considered that the risk of death estimated at time t is independent of the risks estimated at previous times. In addition, the population is considered to be homogeneous, in the sense that the risk distribution is estimated for the entire population taken into account (without taking into account the effects of individual characteristics). The disadvantage is the size of the sample required as well as the estimation of the parameters which is more complicated (Archaux (2005), Boukhetala et al (2009)). If X, a random variable, represents the time elapsed since an instant t_0 and when the time is considered in a discrete manner, if t_i represents an instant during which there is the observation of at least one event, then the probability of survival at time t_i is equal to the probability of having survived before t_i multiplied by the "conditional" probability of surviving at time t_i. The use of the term "conditional" means here that it is about the probability of surviving time t_i knowing that the individuals were survivors in t_i :

$$S(t_i) = P(X > t_i / X \geq t_i) * S(t_{i-1})$$

the probability of survival at t_i then becomes:

$$\begin{aligned} S(t_i) &= S(t_{i-1})(1 - h_i) \\ &= S(t_{i-1}) \frac{n_i - d_i}{n_i} \end{aligned}$$

such as

t_i represents the follow-up time since inclusion in the study for each individual i.

$d_i$is the number of deaths at time t_i.

n_i is the number of subjects at risk of presenting the event studied at the moment t_i , i.e. the number of patients who have not yet undergone the event nor the censorship just before t_i.

By extension, if we consider $t_1 < t_2 < \cdots < t_n$ the distinct survival times of n individuals, $\hat{S}(t)$ corresponds to the product of all the probabilities of not having known the event since the start of observation:

$$\hat{S}(t) = \begin{cases} \prod_{t_i \leq t} \left(1 - \frac{d_i}{N_i}\right) = \prod_{t_i \leq t} (1 - h_i) & \text{si } \; t \geq \; t_1 \\ 1 & \text{si } \; t < t_1 \end{cases} \qquad (1)$$

3. The Bayesian conception of the Kaplan Meier estimator

In the frequentist approach the number of deaths in the interval of time is an realization of a Binomial distribution written by:

$$d_i \sim \beta in(n_i, q_i) \qquad (2)$$

From a Bayesian perspective we assume an a priori for $q_{i,}$, and when the distribution used in the case of proportions is that of Beta, we set:

$$q_{i\sim} beta(\alpha, \beta) \qquad (3)$$

This a priori distribution has several important characteristics in our situation:

- The mean of the beta distribution made it possible to control the precision of the a priori information (informative, non-informative).
- The ease of finding the distribution a posteriori.
- Flexibility of form.
- In the use of Gibbs sampling or in rechanting methods in general, it has a remarkable efficiency because the Gibbs sampler performs a systematic update of each coordinate of the previous state in order to obtain the new state of the chain.

For the hyperparameters (α, β), we use a vague prior distribution, it is a proper distribution with a very large variance, according to this distribution, the prior distribution is considered to be weak informative, and we use this distribution for regularization and stabilization, it provides solutions in the use of algorithms. We pose:

$$q_i \sim \beta(0{,}01{,}0{,}01)$$

For a binomial distribution and a conjugate prior distribution we set

$$f_\pi(d_i/\alpha, \beta) = \int_0^1 f(d_i/q_i)\, \pi(q_i/\alpha, \beta) dq_i$$

$$= \int_0^1 [q_i(1 - q_i)]^{-1} C_{n_i}^{d_i} {q_i}^{d_i} (1 - q_i)^{n_i - d_i} \, dq_i$$

$$= C_{n_i}^{d_i} \frac{1}{B(\alpha, \beta)} \int_0^1 {q_i}^{d_i + \alpha - 1} (1 - q_i)^{n_i - d_i + \beta - 1} \, dq_i$$

$$= C_{n_i}^{d_i} \frac{B(\alpha + d_i, n_i + \beta - d_i)}{B(\alpha, \beta)}$$

which provides a beta – binomial distribution to estimate $\hat{\alpha}, \hat{\beta}$, in order to calculate $\pi\left(q_i / d_i, \hat{\alpha}, \hat{\beta}\right)$. we also pose:

$$\begin{cases} n_1 = the\ number\ of\ subjects\ at\ the\ start\ of\ the\ study \\ n_i = n_{i-1} - d_i - c_i \end{cases}$$

4. the average value of the risk in the case of data of strongly censored durations

In the case of heavily censored duration data, the number of deaths over time is a Binomial distribution given by

$$d_i \sim \beta in(n_i, q_i) \tag{4}$$

and

$$q_{i \sim} beta(\alpha, \beta) \tag{5}$$

in the case of data of heavily censored durations is not obliged to use the hierarchical version of the a priori distribution therefore: $\alpha = \beta = 0.01$.

In this model we assume that $q_{i,}$is constant and follows a saddle distribution

$$\widehat{q}_{moyenne} = \hat{q}_{j,r} \tag{6}$$

we assume the categorial density:

$$r \sim cat(w_i); \sum_{i=1}^{n} w_i = 1$$

we also assume weights for the parameters $w_1, w_2, \dots, w_n$. Such as

$$w_1 + w_2 + \cdots + w_n = 1$$

the parameters $w_1, w_2, \dots, w_n$ represent a function of the censored data having lower weights than those of the uncensored data, then the average risk value is realized via a point approximate approach of the uncensored risks.

We use an arbitrary function that gives influence to weakly censored data: we pose;

$$e_i = \frac{c_i}{\sum_{i=1}^{m} c_i}$$

and

$$w_i = \frac{1 - e_i}{\sum_{i=1}^{m} (1 - e_i)}$$

So the difference between the mean values of the chance risks for two survival functions is:

$$\widehat{q}_{moyenne,a} - \widehat{q}_{moyenne,b} = e - value \qquad (7)$$

e-value is a probability that measures the probable difference or similarity between the average risk of the two survival functions. When we have two hypotheses of the following form:

$$\begin{cases} H_0 & e - value = 0.5 \\ H_1 & e - value \neq 0.5 \end{cases}$$

Il est possible de simplifie les deux hypothèses selon un forme d'un tableau comme suivant :

Table 1: Decision table for the e-value statistic

$\boldsymbol{e - value}$	**Statistical decision**
$e - value > 0.5$	The risk of treatment A is higher than the risk of treatment B
$e - value = 0.5$	Equality between average risks
$e - value < 0.5$	The risk of treatment B is higher than the risk of treatment A

5. Application

Within this segment, a clinical study estimates the survival mechanism of two prescription substances (placebo and prednisolone); this example uses the survival times of 42 patients with chronic active hepatitis. These patients were randomized into two distinct groups; one received treatment with prednisolone and the other received a placebo (see Held, 2010).

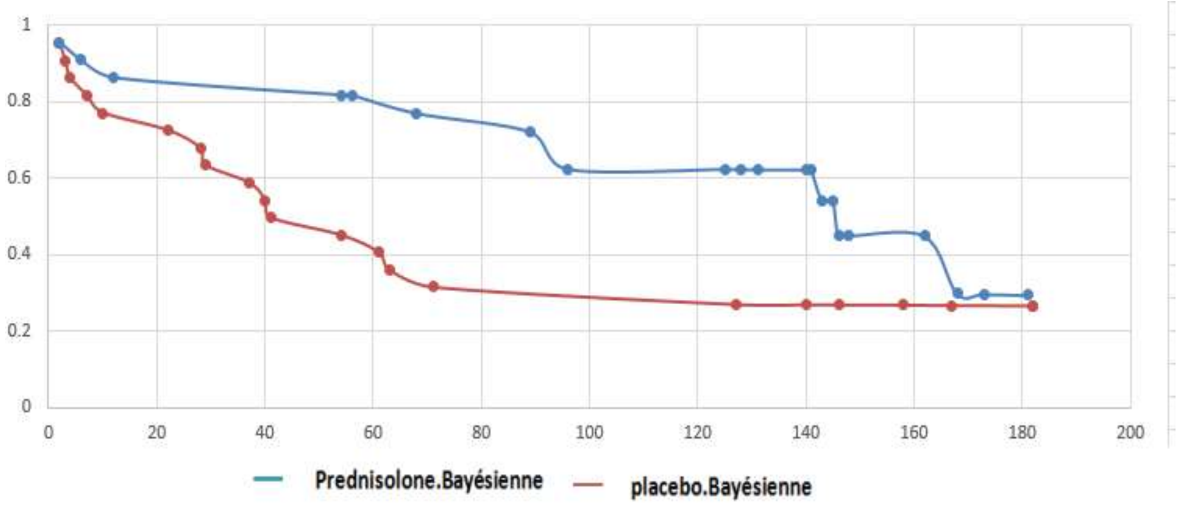
Table 2: Data and survival function for the Prednisolone group.

Time	Total No of Deaths	Total No of censored	No at risk	Prednisolone S(t)
2	1	0	21	0.9545
6	1	0	20	0.9082
12	1	0	19	0.8624
54	1	0	18	0.8169
56	0	1	17	0.8164
68	1	0	16	0.7686
89	1	0	15	0.7198
96	1	0	14	0.6233
125	0	1	13	0.6228
128	0	1	12	0.6223
131	0	1	11	0.6218
140	0	1	10	0.6211
141	0	1	9	0.6204
143	1	0	8	0.5414
145	0	1	7	0.5406
146	1	0	6	0.4502
148	0	1	5	0.4492
162	0	1	4	0.4482
168	1	0	3	0.2985
173	0	1	2	0.2969
181	0	1	1	0.294

Table 3: Data and survival function for the Placebo group.

Time	Total No of Deaths	Total No of censored	Placebo
2	1	0	0.9539
6	1	0	0.9081
12	1	0	0.8614
54	1	0	0.8158
56	0	1	0.7703
68	1	0	0.7248
89	1	0	0.6781
96	1	0	0.6334
125	0	1	0.588
128	0	1	0.5421
131	0	1	0.4965
140	0	1	0.451
141	0	1	0.406
143	1	0	0.3604
145	0	1	0.3149
146	1	0	0.2694
148	0	1	0.2689
162	0	1	0.2683
168	1	0	0.2677
173	0	1	0.2668
181	0	1	0.26

Figure 1: the comparison between the two survival functions.



From the survival curves estimated using the Bayesian Kaplan-Meier method, we find the following results that the divergence between the two functions occurs gradually, which means that the efficacy of prednisolone requires time to appear. Also the survival function of the two pharmaceutical products demonstrates that patients in the placebo group have a lower probability of survival than those in the prednisolon group, and according to this result prednisolone has an efficacy compared to the comparator (placebo). The graph is in the form of "transient difference", where the convergence of the two functions is after the duration 168, which means that prednisolone will lose its effectiveness after this duration.

We now want to go beyond the graphical comparison of the two treatment groups to perform an appropriate statistical test. The Student t test does not lend itself to this, because it asks to ignore that in a part of the patients, the event has not yet taken place. Also the log-rank tests, Wilcoxon and Taron-Ware are developed for a reasonable number of censored data. Therefore, e-value is used to reduce the effect of censorship in the data.

In a first step, we calculate the e-value between the two survival functions of the example without any consideration in terms of censorship. The e-value is a probability and according to table (4) this value is clearly lower than the value 0.5, so the difference between the two survival functions is significant.

Table 4: the test statistic in the absence of the weights of the censored data (See A1).

	mean	**sd**	**MC_error**	**val2.5pc**	**median**	**val97.5pc**
e. value	**0.3814**	0.4857	0.005884	0.0	0.0	1.0

If we can show the study by a comparison, with the classical methods we find:

Table 5: Tests of equality of survival functions.

Statistique	Valeur observée	Valeur critique	p-value	Alpha
Log-Rank	4,028	3,841	0,045	0,05
Wilcoxon	5,686	3,841	0,017	0,05
Tarone -Ware	5,255	3,841	0,022	0,05

From the results summarized in Table (5), the p-value of the log-rank, Wilcoxon and Taron-Ware tests (see Table (1)) is below the threshold (5%), which means that we proved that there is a statistically significant difference between the survival probabilities of the two groups. So the two classical and Bayesian results are identical.

Table 6: the test statistic in the case of using the weights of the censored data (See A2).

	mean	**sd**	**MC_error**	**val2.5pc**	**median**	**val97.5pc**
e. value	**0.3587**	0.4796	0.002044	0.0	0.0	1.0

In Table (6), we calculate the e-value between the two survival functions of the example with the censorship weights. The e-value is clearly less than the value 0.5, so the difference between the two survival functions is significant. The difference also that we find and that between the e-values where there is a reduction of 3% in the probability of difference: i.e. the difference between the two two survival functions increases by 3%, this value is small due to the size of the sample and the number of censored data there are cases where the difference is large.

6. Conclusion

The objective of this article is to test the equality between heavily censored survival functions. We use the comparison between the average risk values or the statistic (e-value). This value calculated in the Kaplan Meier model and according to a Bayesian design and through the posterior mean approach. In the results we find:

- the. value is the probability of equality of the mean risks of the chance functions, this probability changes in the case of considering the weight of the censored data.
- the. value gives credibility to the results found because the calculation of the risk average is not identical for all durations because of incomplete data or censored data.

7. Référence

- Archaux, C (2005), Conception d'un système d'information dédie à l'estimation de la valeur des clients en téléphonie mobile prépayée. Thèse de doctorat, Ecole Nationale Supérieure des Ingénieurs d'Etudes et Techniques de l'Armement de Brest, France.
- Barnard GA (1958), Thomas Bayes - a biographica1 note. Biomelrika 45: 293-315.
- Boukhetala, K., Marion, J.M., Oulidi, A (2009), Apport des méthodes de durée de vie au domaine de l'assurance. Publication dans la 41émes journées de statistique, Bordeaux.
- Carlin, B.P., Gelfand, A.E.. Smith, A.F.M (1992), Hierarchical Bayesian Ana1ysis of Changepoint Problems. Appl. Slalisl .. 41, No.2, 389-405.
- Feigl, P. et Zelen, M (1965), Estimation of exponential survival probabilities with concomitant information Biome· trics 21 : 826-838.
- Ferguson, T.S (1973), A Bayesian analysis of sorne nonparametric problems. The Annals of Statistics 1 :209-230.
- Held, U (2010), Représentation graphique et comparaison de courbes de survie. Forum Med Suisse, 10(33), 548-550.
- Kalbfleisch, J. D (1978), Non-parametric Bayesian analysis of survival time data. Journal of the Royal Statistical Society: Series B (Methodological), 40(2), 214-221.
- Kaplan, E.L., Meier, P (1958), Non-parametric estimation from in complete observations. Journal of the American Slatisticai Associalion 53: 457-481.

8. Appendices (OpenBUGS codes)

A1.

```
model
{
for (i in 1:m1) {
d1[i]~dbin(q1[i],n1[i])
q1[i]~dbeta(0.01,0.01)
w1[i]<-1/m1
}
for (i in 1:m2) {
d2[i]~dbin(q2[i],n2[i])
q2[i]~dbeta(0.01,0.01)
w2[i]<-1/m2
}
for (i in 1:m1){
ce1[i]~dbin(0.01,0.01)
}
for (i in 1:m2){
ce2[i]~dbin(0.01,0.01)
}
for (i in 1:m1){
qc1[i]~dbeta(0.01,0.01)
}
for (i in 1:m2){
qc2[i]~dbeta(0.01,0.01)
}
for (i in 1:m1){
p1[i]<-1-q1[i]
}
for (i in 1:m2){
p2[i]<-1-q2[i]
}
n1[1]<- 22
n2[1]<- 22
for(i in 2:m1){
n1[i]<-n1[i-1]-d1[i-1]-ce1[i-1]
}
for(i in 2:m2){
n2[i]<-n2[i-1]-d2[i-1]-ce2[i-1]
}
for (i in 2:m1){
s1[i]<-s1[i-1]*p1[i]
 }
s1[1]<-p1[1]
for (i in 2:m2){
s2[i]<-s2[i-1]*p2[i]
 }
s2[1]<-p2[1]
q.avg1 <- q1[r] #Composite posterior, monitor this node
r ~ dcat(w1[])
q.avg2 <- q2[r2] #Composite posterior, monitor this node
r2 ~ dcat(w2[])
e.value <- step(q.avg1 - q.avg2)
}
list(m1=21,d1=c(1,1,1,1,0,1,1,1,0,0,0,0,0,1,0,1,0,0,1,0,0),ce1=c(0,0,0,0,1,0,0,0,1
,1,1,1,1,0,1,0,1,1,0,1,1),m2=22,d2=c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0),c
e2=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1))
```

A2.

```
model
{
for (i in 1:m1) {
d1[i]~dbin(q1[i],n1[i])
q1[i]~dbeta(0.01,0.01)
}
for (i in 1:m2) {
d2[i]~dbin(q2[i],n2[i])
q2[i]~dbeta(0.01,0.01)
}
for (i in 1:m1){
ce1[i]~dbin(0.01,0.01)
}
for (i in 1:m2){
ce2[i]~dbin(0.01,0.01)
}
for (i in 1:m1){
qc1[i]~dbeta(0.01,0.01)
}
for (i in 1:m2){
qc2[i]~dbeta(0.01,0.01)
}
for (i in 1:m1){
p1[i]<-1-q1[i]
}
for (i in 1:m2){
p2[i]<-1-q2[i]
}
n1[1]<- 22
n2[1]<- 22
for(i in 2:m1){
n1[i]<-n1[i-1]-d1[i-1]-ce1[i-1]
}
for(i in 2:m2){
n2[i]<-n2[i-1]-d2[i-1]-ce2[i-1]
}
for (i in 2:m1){
s1[i]<-s1[i-1]*p1[i]
 }
s1[1]<-p1[1]
for (i in 2:m2){
s2[i]<-s2[i-1]*p2[i]
 }
s2[1]<-p2[1]
q.avg1 <- q1[r] #Composite posterior, monitor this node
r ~ dcat(w1[])
q.avg2 <- q2[r2] #Composite posterior, monitor this node
r2 ~ dcat(w2[])
e.value <- step(q.avg1 - q.avg2)
 }
```

```
list(m1=21,d1=c(1,1,1,1,0,1,1,1,0,0,0,0,0,1,0,1,0,0,1,0,0),ce1=c(0,0,0,0,1,0,0,0,1
,1,1,1,1,0,1,0,1,1,0,1,1),m2=22,
d2=c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0),ce2=c(0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1),
w1=c(0.05,0.05,0.05,0.05,0.045454545,0.05,0.05,
0.05,0.045454545,0.045454545,0.045454545,0.045454545,0.045454545,0.05,0.
045454545,
0.05,0.045454545,0.045454545,0.05,0.045454545,0.045454545),w2=c(0.05,0.0
5,0.05,0.05,0.05,0.05,0.05,0.05,0.05,0.05,0.05,0.05,0.05,0.05,0.05,0.041666667,
0.041666667,0.041666667,0.041666667,0.041666667,0.041666667))
```