

The posterior mean approach to determine the mean value of risk in the case of heavily and weakly censored data

HAMIMES Ahmed *	BENAMIROUCHE Rachid
Assistant professor "A"	Professor
ahmedhamimes@yahoo.com	rbena2002@hotmail.com
Faculty of Medicine, University of Constantine 3. Algeria.	National School of Statistics and Applied Economics. Alegria.

Received in: 11/10/2020

accepted in: 01/11/2020

published in: 31/12/2020

Summary

In this paper, Kaplan Meier's model is used in the survival analysis and according to a Bayesian conception of the context to calculate the mean value of the risk in the case of data of durations strongly and weakly censored through the approach of the posterior mean. This method makes it possible to probabilize the mean value of the risk of chance and to make comparisons in the same way.

Keywords: the posterior mean approach, Kaplan Meier model, survival analysis, strongly and weakly censored data.

JEL Classification Codes : C11, C12, C41.

Résumé

Dans cette article, on utilise le modèle de Kaplan Meier dans l'analyse de survie et selon un conception bayésien de le contexte de calculer la valeur moyenne du risque dans le cas de données de durées fortement et faiblement censurées à travers l'approche de la moyenne a posteriori (The posterior mean approach). Cette méthode permet de probabilisée la valeur moyenne du risque de hasard et de faire des comparaisons de la même façon.

Mots clés : l'approche de la moyenne a posteriori, le modèle de Kaplan Meier, modèles de durées, données fortement et faiblement censurées.

JEL Classification Codes : C11, C12, C41.

* Corresponding author

Introduction

Survival analysis refers to a set of statistical techniques used to process censored data, that is, data for which, for some of them, only an upper or lower limit and not a precise value. . The application of survival models can be found in several disciplines, for example, survival analysis is used in medicine to assess the effectiveness of a treatment. For example, we want to estimate the probable survival time of a patient, for this we use a list of patients that we know, for each,

- either the actual survival time (data not filtered or detected),
- or a lower limit (censored data) for this period.

The second case occurs when, for example, a patient is lost due to movement or dies from an independent cause. In demography, the measure of survival is used to construct life tables. These are used by actuaries, among others, to assess the value of life insurance and annuities; when the data is divided into intervals, we speak of actuarial tables; Engineering survival analysis is used to assess the efficiency of electronic devices and components. Analysis of survival is also useful in astrophysics. The Kaplan-Meier method can quickly obtain a survival curve, as well as essential statistics such as the median residual survival time.

In the case of semi-parametric models, several authors have discussed the use of Bayesian inference in survival analysis (Ferguson 1973; Kalbfleisch 1978; Kalbfleisch and Prentice 1980; Clayton 1978) in which the a priori of Base rates and regression coefficients of covariates are specified in different ways. Some of these Bayesian methods use total probability rather than partial probability. Indeed, one of the advantages of using Bayesian approaches to jointly model the regression coefficients of the covariates and the base death rate is that by using MCMC techniques, one can precisely quantify the a posteriori of the model as well as their standard deviations. The functions of specification of a priori rational distribution as well as the execution of intensive computations remain. The results of these Bayesian proportional hazard

experiments demonstrated the precision of the calculations and the possible benefits of using these approaches to assess survival data (Giorgi, 2002).

Parametric models have taken a significant place in the study of Bayesian survival, parametric modeling provides direct modeling. The statistical literature in Bayesian parametric approach in analyzes and survival tests is very extensive, we give here some references in the medical or public health field (Grieve, 1987; Achcar et al., 1987; Achcar et al., 1985 ; Chen et al., 1985; Dellaportas and Smith, 1993; Kim and Ibrahim, 2001), the book by Ibrahim et al. (2001) provides a clear description of Bayesian survival models in general, and parametric models in particular. We can also cite the Bayesian method for breaking point models introduced by Carlin, Gelfand and Smith (1992) in parametric modeling.

Florens and Rolin (2001) mainly demonstrated in nonparametric modeling that the simulation estimates of the Dirichlet process and nonparametric Bayesian inference give good results compared to classical methods. These different works have been developed in different methodological contexts, using specific a priori distributions and / or modeling the role of cumulative risk or directly the role of instantaneous risk (Giorgi 2002). Different parametric, semi-parametric and non-parametric models were built according to the classic method, (quote here, K-M, Gamma, Log-logistic, Cox (classic version) ...

The Kaplan-Meier method makes it possible to estimate the survival functions, without requiring regular time intervals, unlike the actuarial life table method. Survival curves are used over time to analyze changes in the size of a given population. The average value of the risk provided is probabilized. We therefore need a Bayesian approach that preserves this variability from one source to another. Such an analysis is called a hierarchical Bayesian analysis because steps or hierarchies are used to establish the prior distribution.

We first specify a prior to represent the variability of the data sources (the data should not be pooled (see Dezfuli et al (2009)), then we specify a second a priori describing the epistemic uncertainty in the parameters of the a priori first step. The analysis is quite complicated mathematically, but WinBUGS makes the analysis simple. In the result we find the average variability curve of the beta population (Average beta population variability curve). Consequently This method allows to probabilized the mean value of the chance risk and make comparisons in the same way.

1. Kaplan Meier model

The Kaplan-Meier estimator (1958) is a functional method for estimating the survival function is written

$$\hat{S}(t) = \begin{cases} \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) = \prod_{t_i \leq t} (1 - q_i) = \prod_{t_i \leq t} p_i & \text{si } t \geq t_1 \\ 1 & \text{si } t < t_1 \end{cases} \quad (1)$$

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 .

2. The Bayesian conception of the Kaplan Meier estimator

In a Bayesian view it is assumed that the number of deaths in the interval of time is a Binomial distribution given by

$$d_i \sim \text{bin}(n_i, q_i) \quad (2)$$

In this discretized approach of the Kaplan Meier model, the risk of death estimated by the likelihood method is

$$\hat{q}_i = \frac{d_i}{n_i},$$

The parameters q_i , in the Bayesian framework are random variables, and when the distribution used in the case of proportions is that of Beta, we set:

$$q_i \sim \text{beta}(\alpha, \beta) \quad (3)$$

The prior distribution is considered to be weak informative, it provides solutions in the use of algorithms. We pose:

$$q_i \sim \beta(0.01, 0.01) \quad (4)$$

In this article we use the hierarchical method in the construction of a priori distributions, then

$$q_i \sim \beta(\alpha, \beta) \quad (5)$$

and

$$\alpha \sim \text{Gamma}(0.0001, 0.0001) \quad (6)$$

$$\beta \sim \text{Gamma}(0.0001, 0.0001) \quad (7)$$

3. the average value of the risk in the case of data of weakly censored durations

the number of deaths in the interval of time is a Binomial distribution given by

$$d_i \sim \text{bin}(n_i, q_i) \quad (8)$$

and

$$q_i \sim \text{beta}(\alpha, \beta) \quad (9)$$

In this step we use the hierarchical method in the construction of the a priori distributions, then

$$q_i \sim \beta(\alpha, \beta) \quad (10)$$

and

$$\alpha \sim \text{Gamma}(0.0001, 0.0001) \quad (11)$$

$$\beta \sim \text{Gamma}(0.0001, 0.0001) \quad (12)$$

In this model we assume that q_i is constant and follows a saddle distribute

$$q_{\text{moyenne}} \sim \text{beta}(\alpha, \beta) \quad (13)$$

To check the quality of the Kaplan Meier Bayesian model in this operation we use the observed values of d_i to form the statistic

$$\chi_{obs}^2 = \sum_i \frac{(d_{obs,i} - \mu_i)^2}{\sigma_i^2} = \sum_i \frac{(d_{obs,i} - (q_i * n_i))^2}{(q_i * n_i) * (1 - (q_i))} \quad (14)$$

We then generate predicted values of d_i from its posterior predictive distribution, and construct an analogous statistic:

$$\chi_{resp}^2 = \sum_i \frac{(d_{resp,i} - \mu_i)^2}{\sigma_i^2} = \sum_i \frac{(d_{resp,i} - (q_i * n_i))^2}{(q_i * n_i) * (1 - (q_i))} \quad (15)$$

evaluating the posterior distributions of $D(d_{obs,i}, q_i)$ and $D(d_{resp,i}, q_i)$ provides individual and aggregated measures of goodness of fit that can be described graphically or in using tail region probabilities called a posteriori predictive values (Meng, 1994).

$$p - value \equiv P[D(d_{resp,i}, q_i) \geq D(d_{obs,i}, q_i)/d_i]$$

Gelman, Meng and Stern (1996) recommend calculating the "predictive p-value"

$$p - value \equiv P[D(d_{resp,i}, q_i) \geq D(d_{obs,i}, q_i)/d_i]$$

$$= \int \int I_{[D(d_{resp,i}, q_{constant}) \geq D(d_{obs,i}, q_{constant})]} f(d_i^{resp}/q) \pi(q_i/d_i) d_i^{resp} d\theta \quad (16)$$

where I is the indicator function.

This integral can be approximated by sampling q_i^k from the posterior distribution of q_i , the same thing for d_i^{resp} rep from the distribution $f(d_i^{resp}/q_i)$. In the result we find

$$p - value = \sum_{k=1}^T I[D(d_i^k, q_i^k) \geq D(d_i, q_i^k)]/T \quad (17)$$

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

In the case of data of heavily censored durations, the number of deaths in the time interval is a Binomial distribution given by

$$d_i \sim \beta in(n_i, q_i) \quad (18)$$

and

$$q_i \sim \text{beta}(\alpha, \beta) \quad (19)$$

In the case of data of strongly 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

$$\hat{q}_{moyenne} = \hat{q}_{j,r} \quad (13)$$

we assume the categorial density:

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

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

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

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

5. Applications

5.1. Application 1 (the average value of the risk in the case of data of weakly censored durations)

In this section, survival function is estimated in a clinical study for two pharmaceuticals (placebo and prednisolone), this example uses survival times for 42 patients with chronic active hepatitis. These patients were randomized to two equal groups, one was treated with prednisolone, the other received a placebo (see Held, 2010). In this example, patients with prednisolone are used.

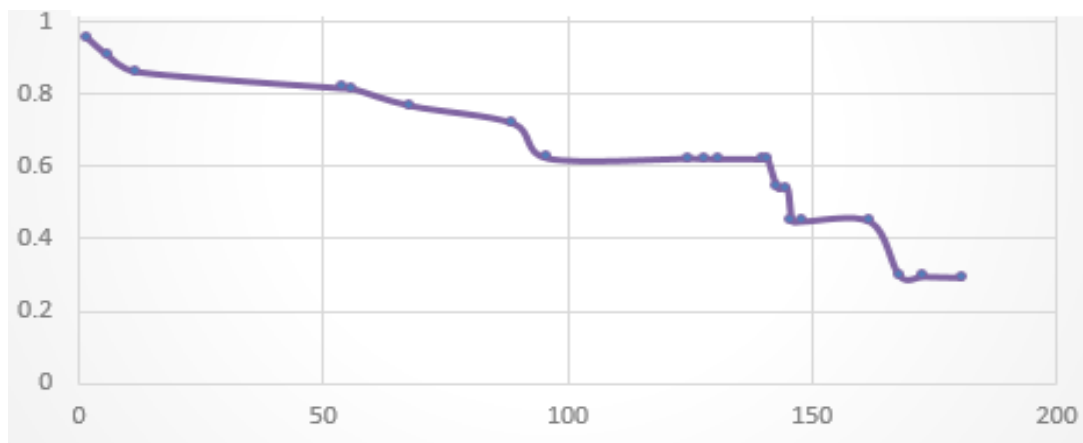
Table 1: Survival probabilities by the Bayesian Kaplan Meier method.

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

Source: Held, 2010.

If we assume the absence of any a priori information on the estimated survival model, the choice of an uninformative a priori is obvious. We use in this article a conjugate prior distribution such that $\alpha = \beta = 0.01$.

Figure 1- The survival curve estimated according to the Bayesian Kaplan-Meier method with a priori of beta.



Source: Developed by us using Excel.

From Figure (1), we notice that at the start of the curve, 100% of the individuals in the sample are included in the treatment study. After more than 146 days after using the

treatment in the sample 50% of the patients had died. But, the treatment failure for the rest of the individuals in the sample lasts for a long time, for some it exceeds 180 days.

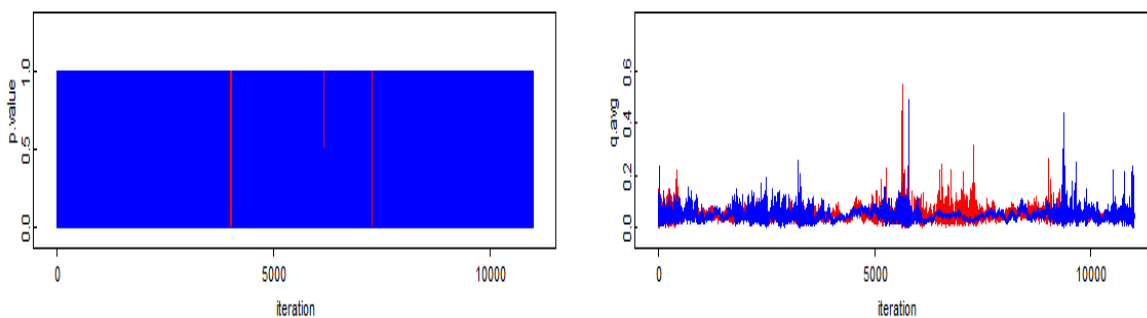
Table 2: The parameters of the Kaplan Meier model estimated.

	mean	sd	MC_error	val2.5pc	median	val97.5pc
p.value	0.555	0.497	0.01294	0.0	1.0	1.0
q.avg	0.04598	0.01977	7.345E-4	0.01468	0.04397	0.09007

Source: Developed by us using Openbugs.

The a posteriori predictive p-value (p-value = 0.555) can be directly interpreted as the probability of observing in future samples with $D(d_i, q_i^k)$ higher than that already observed. This value is close to 0.5 so the repeated and real data distributions are close, while values close to zero or one indicate differences between them (Gelman and Meng, 1996). The Hierarchical Bayesian Kaplan Meier model effectively represents the sample. The value of q.avg indicates that the average chance of dying during a day when taking prednisolone is 4%.

Figure2: The trace of the posterior distribution for the model parameters.



Source: Developed by us.

In Figure 2, each color denotes an MCMC chain. The two chains mix well: convergence is achieved (see A1 in the appendix).

5.2. Application 2 (the average value of the risk in the case of data of strongly censored durations)

For the proposed approach we use the study of the real data set "Stanford Heart Transplant results" (Kalbflesch and Printice, 1980), which is the classic data set for survival. If we assume that $\alpha = \beta = \alpha' = \beta' = 0.01$, we find the following table:

Table 3: Survival probabilities by the Bayesian Kaplan Meier method of the example

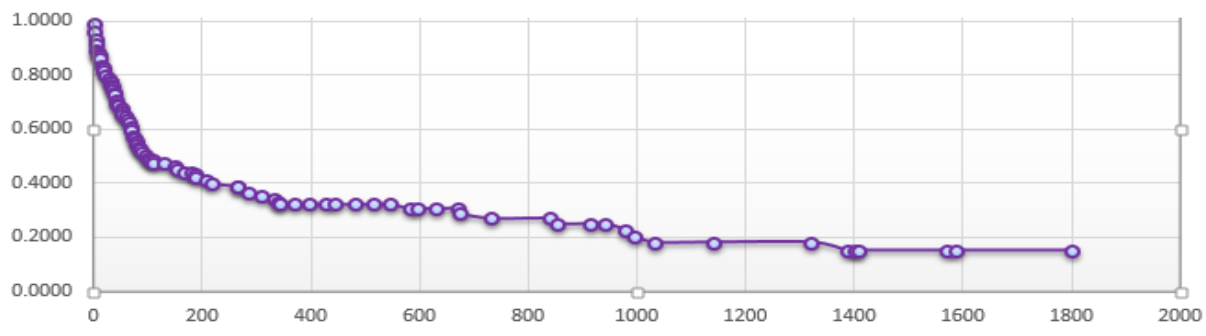
Time	Total No of Deaths	Total No of censored	No at risk	Kaplan Meier	Time	Total No of Deaths	Total No of censored	No at risk	Kaplan Meier
1	1	0	103	0.9903	131	0	1	46	0.4736
2	3	0	102	0.9612	149	1	0	45	0.4631
3	3	0	99	0.932	153	1	0	44	0.4526
5	2	0	96	0.9126	165	1	0	43	0.442
6	2	0	94	0.8932	180	0	1	42	0.442
8	1	0	92	0.8835	186	1	0	41	0.4313
9	1	0	91	0.8738	188	1	0	40	0.4205
11	0	1	90	0.8738	207	1	0	39	0.4097
12	1	0	89	0.864	219	1	0	38	0.3989
16	3	0	88	0.8345	263	1	0	37	0.3881
17	1	0	85	0.8247	265	0	1	36	0.3881
18	1	0	84	0.8149	285	2	0	35	0.366
21	2	0	83	0.7952	308	1	0	33	0.3549
28	1	0	81	0.7854	334	1	0	32	0.3438
30	1	0	80	0.7756	340	1	1	31	0.3327
31	0	1	79	0.7756	342	1	0	29	0.3212
32	1	0	78	0.7657	370	0	1	28	0.3212
35	1	0	77	0.7557	397	0	1	27	0.3212
36	1	0	76	0.7458	427	0	1	26	0.3212
37	1	0	75	0.7358	445	0	1	25	0.3212
39	1	1	74	0.7259	482	0	1	24	0.3212
40	2	0	72	0.7057	515	0	1	23	0.3212
43	1	0	70	0.6956	545	0	1	22	0.3212
45	1	0	69	0.6856	583	1	0	21	0.3059
50	1	0	68	0.6755	596	0	1	20	0.3059
51	1	0	67	0.6654	630	0	1	19	0.3059
53	1	0	66	0.6553	670	0	1	18	0.3059
58	1	0	65	0.6452	675	1	0	17	0.2879
61	1	0	64	0.6352	733	1	0	16	0.2699
66	1	0	63	0.6251	841	0	1	15	0.2699
68	2	0	62	0.6049	852	1	0	14	0.2507
69	1	0	60	0.5948	915	0	1	13	0.2507
72	2	0	59	0.5747	941	0	1	12	0.2507

77	1	0	57	0.5646	979	1	0	11	0.2279
78	1	0	56	0.5545	995	1	0	10	0.2051
80	1	0	55	0.5444	1032	1	0	9	0.1823
81	1	0	54	0.5343	1141	0	1	8	0.1823
85	1	0	53	0.5243	1321	0	1	7	0.1823
90	1	0	52	0.5142	1386	1	0	6	0.1519
96	1	0	51	0.5041	1400	0	1	5	0.1519
100	1	0	50	0.494	1407	0	1	4	0.1519
102	1	0	49	0.4839	1571	0	1	3	0.1519
109	0	1	48	0.4839	1586	0	1	2	0.1519
110	1	0	47	0.4736	1799	0	1	1	0.1519

Source: Developed by us.

Figure.1- The survival curve estimated according to the Kaplan-Meier method of example

(1).



Source: Developed by us using OpenBUGS.

From Figure (1), it can be seen that at the start of the curve, 100% of the individuals in the sample are included in the "Stanford Heart Transplant results" study. After approximately 96 days after heart transplantation at this hospital 50% of patients had died. But heart transplant failure for the rest of the individuals in the sample lasts a long time, some even exceeding four years.

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

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

and

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

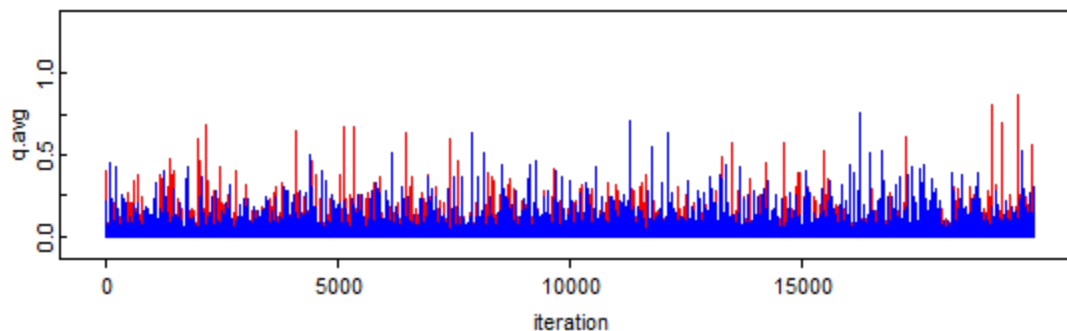
Table 3: Bayesian estimation of the parameters q.avg

	mean	sd	MC_error	val2.5pc	median	val97.5pc
q.avg	0.02108	0.04071	2.148E-4	0.0	0.008049	0.1247

Source: Developed by us OpenBUGS.

The value of q.avg indicates that the average probability of death in one day after a Stanford heart transplant is 2% considering the weights of the censored data.

Figure 3 : The trace of the posterior distribution for the parameter q.avg.



Source: Developed by us using OpenBUGS.

In Figure 3, each color denotes an MCMC chain. The two chains mix well: convergence is achieved (See A2 in the appendix).

Conclusion

In this paper, Kaplan Meier's Bayesian model was used in the survival analysis with the objective of calculating the mean value of risk in the case of strongly and weakly censored duration data using the mean approach a. posteriori (The posterior mean approach) and in two different examples, in the result we find that this method makes it possible to verify the validity of the Bayesian model of Kaplan Meier.

References

1. DEZFULI, H., KELLY, D., SMITH, C., VEDROS, K., & GALYEAN, W(2009), *Bayesian inference for NASA probabilistic risk and reliability analysis*;
2. FERGUSON, T.S., (1973), *A Bayesian analysis of some nonparametric problems*, The Annals of Statistics, 1(2), 209-230;
3. FLORENS, J.P., ROLIN, J.M.,(2001), *Simulation of posterior distributions in nonparametric censored analysis*, International Statistical Review, 67(2), 187-210;
4. CLAYTON, D.G.,(1978), *A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence*, Biometrika, 65, 141 – 151;
5. KALBFLEISCH, J.D.,(1978), *Nonparametric bayesian analysis of survival time data*, Journal of the Royal Statistical Society, Series B, 40(2), 214-221;
6. KALBFLEISCH, J.D., PRENTICE, R.I., (1980), *The Statistical analysis of failure time data*, John Wiley & Sons, New York;
7. GIORGI, R.,(2002), *Analyses comparatives des méthodes de survie et extensions d'un modèle régressif de survie relative : prise en compte de la non-proportionnalité des risques par des fonctions B-splines et développement d'une méthode d'analyse bayésienne*, Thèse de doctorat, Université Aix-Marseille II, France;
8. GRIEVE, A.P., (1987), *Applications of Bayesian software: Two examples*, The Statistician, 36, 283-288,;
9. ACHCAR, J.A., BOLFARINE, H., PERICCHI, L.R.,(1986), *Transformation of survival data to an extreme value distribution*, the statistician, 36, 229-234;
10. ACHCAR, J.A., BROOKMEYER, R., HUNTER, W.G.,(1985), *An application of Bayesian analysis to medical follow-up data*, Statistics in Medicine, 4, 509-520 ;
11. CHEN, W.C., HILL, B.M., GREENHOUSE, J.B., FAYOS, J.V., (1985), *Bayesian analysis of survival curves for cancer patients following treatment*, In Bayesian Statistics 2 (Eds. J.O. Berger, J. Bernardo. A.F.M. Smith), Amsterdam: North-Holland, 299-328;
12. DELLAPORTAS, P., SMITH, A.F.M.,(1993), *Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling*, Applied Statistics, 42, 443-459;
13. KIM, S.W., IBRAHIM, J.G.,(2001), *On Bayesian inference for proportional hazards models using noninformative priors*, Lifetime Data Analysis, to appear ;
14. IBRAHIM, J. G., CHEN, M, H., SINHA, D., (2001), *Bayesian survival analysis*, Springer.
15. CARLIN, B.P., GELFAND, A.E., SMITH, A.F.M., (1992), *Hierarchical Bayesian Analysis of Change point Problems*, Appl. Statist. 41(2), 389-405;
16. Held, U. (2010). *Représentation graphique et comparaison de courbes de survie*. Forum Med Suisse, 10(33), 548-550.

• Appendices (OpenBUGS code)

```
Code A1
model {
for (i in 1 : m) {
d[i] ~ dbin(q[i], n[i]) #Binomial model for d
q[i] ~ dbeta(alpha, beta)
```

